

# Deep Generative Models

## Lecture supplementary

Roman Isachenko



Spring, 2022

# Outline I

## 1. Autoregressive models

Masked Autoencoder (MADE)

GatedPixelCNN

## 2. ELBO gradient, Log derivative trick

## 3. Mean field approximation

## 4. IWAE

## 5. PixelVAE, Hierarchical VAE

## 6. Posterior collapse

## 7. Flows intuition

## 8. Parallel WaveNet

## 9. RevNet, i-RevNet

## 10. ELBO surgery

## 11. Disentanglement

InfoGAN

$\beta$ -VAE, Disentanglement metric

FactorVAE

## Outline II

### Challenging Disentanglement Assumptions

#### 12. GANs

DCGAN

Improved techniques for training GANs

WGAN

StyleGAN

#### 13. FFJORD

#### 14. Vector Quantized VAE-2

#### 15. Feature Quantized GAN

C

# Outline

## 1. Autoregressive models

Masked Autoencoder (MADE)  
GatedPixelCNN

## 2. ELBO gradient, Log derivative trick

## 3. Mean field approximation

## 4. IVAE

## 5. PixelVAE, Hierarchical VAE

## 6. Posterior collapse

## 7. Flows intuition

## 8. Parallel WaveNet

## 9. RevNet, i-RevNet

## 10. ELBO surgery

## 11. Disentanglement

## 12. GANs

## 13. FFJORD

## 14. Vector Quantized VAE-2

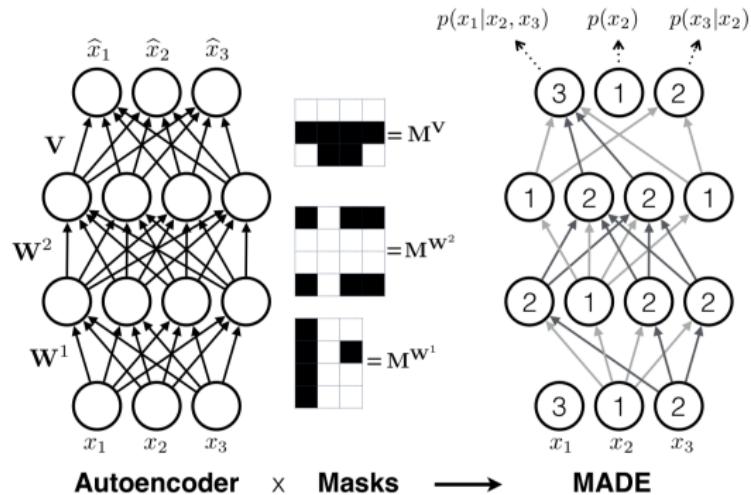
## 15. Feature Quantized GAN

# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

# MADE

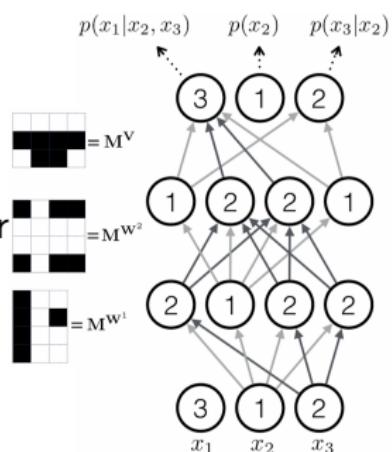
- ▶ Vanila autoencoder is not a generative model.
- ▶ Let mask the weight matrices to make the model generative:  
 $\mathbf{W}_M = \mathbf{W} \cdot \mathbf{M}$ .



- ▶ The question is how to create matrices  $\mathbf{M}$  which produce the autoregressive property?

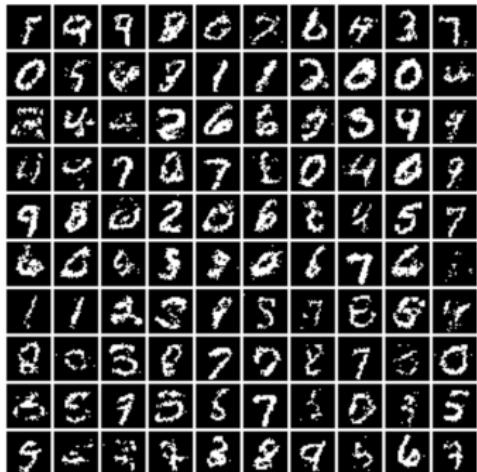
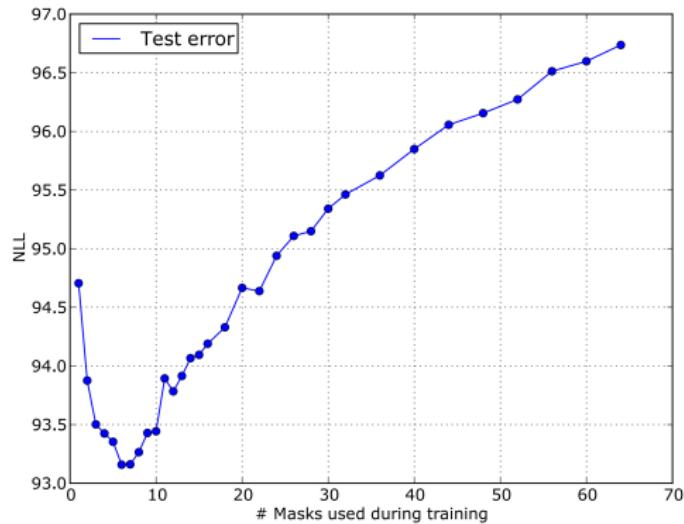
## Masks generation

- ▶ Define the ordering of input elements from 1 to  $m$ .
- ▶ Assign the random number  $k$  from 1 to  $m - 1$  to each hidden unit. The number gives the maximum value of input units to which the unit can be connected.
- ▶ Connect each hidden unit with number  $k$  with the previous layer units which has the number is **less or equal** than  $k$ .
- ▶ Connect each output unit with number  $k$  with the previous layer units which has the number is **less** than  $k$ .



## Possible variations

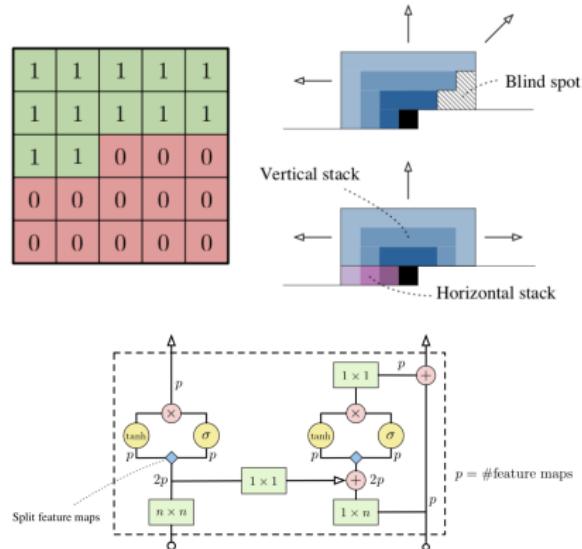
- ▶ Order agnostic training (missing values in partially observed input vectors can be imputed efficiently);
- ▶ Connectivity-agnostic training (cheap ensembling).



# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

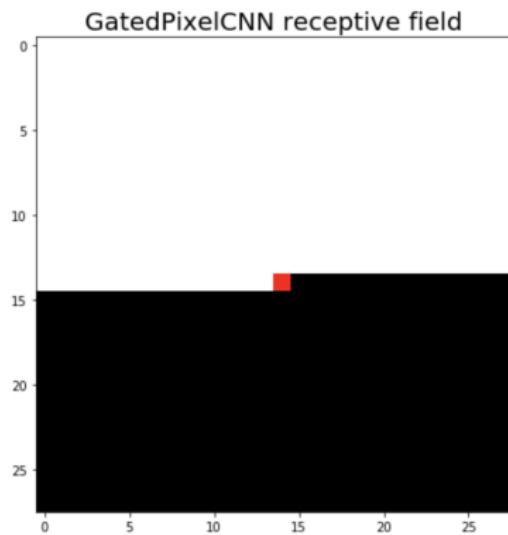
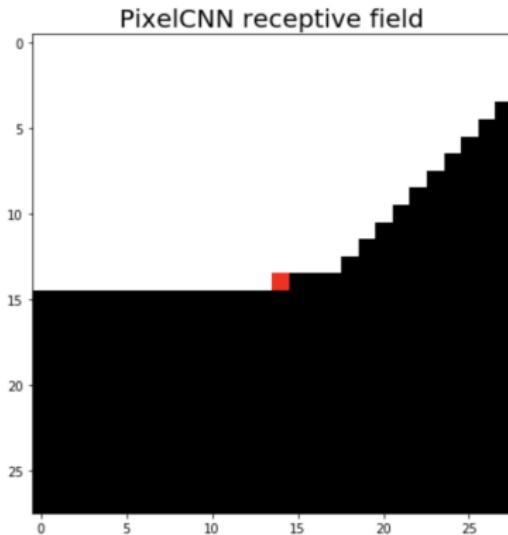
# GatedPixelCNN (2016)



Van den Oord A. et al. Conditional image generation with pixelcnn decoders

<https://arxiv.org/pdf/1606.05328.pdf>

# GatedPixelCNN (2016)



---

Van den Oord A. et al. Conditional image generation with pixelcnn decoders

<https://arxiv.org/pdf/1606.05328.pdf>

## Extensions

- ▶ **PixelCNN++**: *Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications*  
<https://arxiv.org/pdf/1701.05517.pdf>  
(mixture of logistics instead of softmax);
- ▶ **PixelSNAIL**: *An Improved Autoregressive Generative Model*  
<https://arxiv.org/pdf/1712.09763.pdf>  
(self-attention to learn optimal autoregression ordering).

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

## ELBO gradient (E-step, $\nabla_{\phi}\mathcal{L}(\phi, \theta)$ )

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_q \log p(\mathbf{X}|\mathbf{Z}, \theta) - KL(q(\mathbf{Z}|\mathbf{X}, \phi) || p(\mathbf{Z})) \rightarrow \max_{\phi, \theta} .$$

Difference from M-step: density function  $q(\mathbf{z}|\mathbf{x}, \phi)$  depends on the parameters  $\phi$ , it is impossible to use Monte-Carlo estimation:

$$\nabla_{\phi}\mathcal{L}(\phi, \theta) = \int \nabla_{\phi} q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \theta) d\mathbf{Z} - \nabla_{\phi} KL$$

### Log-derivative trick

$$\nabla_{\xi} q(\eta|\xi) = q(\eta|\xi) \left( \frac{\nabla_{\xi} q(\eta|\xi)}{q(\eta|\xi)} \right) = q(\eta|\xi) \nabla_{\xi} \log q(\eta|\xi).$$

$$\nabla_{\phi} q(\mathbf{Z}|\mathbf{X}, \phi) = q(\mathbf{Z}|\mathbf{X}, \phi) \nabla_{\phi} \log q(\mathbf{Z}|\mathbf{X}, \phi).$$

## ELBO gradient (E-step, $\nabla_{\phi}\mathcal{L}(\phi, \theta)$ )

$$\begin{aligned}\nabla_{\phi}\mathcal{L}(\phi, \theta) &= \int \nabla_{\phi}q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \theta) d\mathbf{Z} - \nabla_{\phi}KL = \\ &= \int q(\mathbf{Z}|\mathbf{X}, \phi) [\nabla_{\phi} \log q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \theta)] d\mathbf{Z} - \nabla_{\phi}KL\end{aligned}$$

After applying log-reparametrization trick, we are able to use Monte-Carlo estimation:

$$\begin{aligned}\nabla_{\phi}\mathcal{L}(\phi, \theta) &\approx n \nabla_{\phi} \log q(\mathbf{z}_i^* | \mathbf{x}_i, \phi) \log p(\mathbf{x}_i | \mathbf{z}_i^*, \theta) - \nabla_{\phi}KL, \\ \mathbf{z}_i^* &\sim q(\mathbf{z}_i | \mathbf{x}_i, \phi).\end{aligned}$$

### Problem

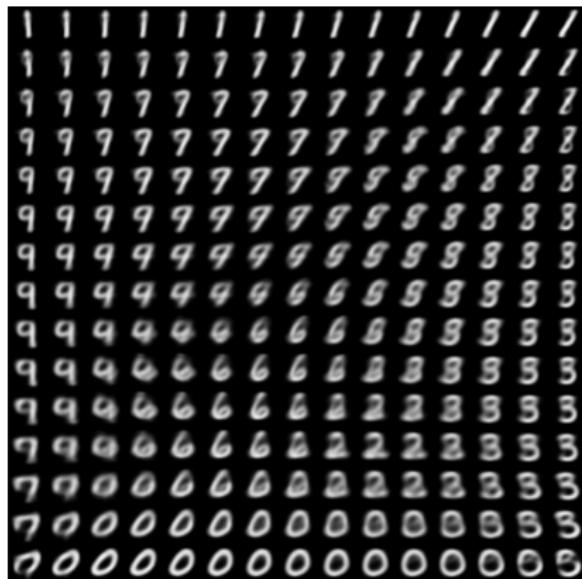
Unstable solution with huge variance.

### Solution

Reparametrization trick

# Variational Autoencoder

Generated images for latent objects  $\mathbf{z}$  sampled from prior  $\mathcal{N}(0, \mathbf{I})$



# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

# Bayesian framework

## Bayes theorem

$$p(\mathbf{t}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{t})p(\mathbf{t})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathbf{t})p(\mathbf{t})}{\int p(\mathbf{x}|\mathbf{t})p(\mathbf{t})d\mathbf{t}}$$

- ▶  $\mathbf{x}$  – observed variables,  $\mathbf{t}$  – unobserved variables (latent variables/parameters);
- ▶  $p(\mathbf{x}|\mathbf{t})$  – likelihood;
- ▶  $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{t})p(\mathbf{t})d\mathbf{t}$  – evidence;
- ▶  $p(\mathbf{t})$  – prior distribution,  $p(\mathbf{t}|\mathbf{x})$  – posterior distribution.

## Meaning

We have unobserved variables  $\mathbf{t}$  and some prior knowledge about them  $p(\mathbf{t})$ . Then, the data  $\mathbf{x}$  has been observed. Posterior distribution  $p(\mathbf{t}|\mathbf{x})$  summarizes the knowledge after the observations.

## Variational Lower Bound

We have set of objects  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ . The goal is to perform Bayesian inference on the unobserved variables  $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^n$ .

### Evidence Lower Bound (ELBO)

$$\begin{aligned}\log p(\mathbf{X}) &= \log \frac{p(\mathbf{X}, \mathbf{T})}{p(\mathbf{T}|\mathbf{X})} = \\&= \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})}{p(\mathbf{T}|\mathbf{X})} d\mathbf{T} = \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})q(\mathbf{T})}{p(\mathbf{T}|\mathbf{X})q(\mathbf{T})} d\mathbf{T} = \\&= \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})}{q(\mathbf{T})} d\mathbf{T} + \int q(\mathbf{T}) \log \frac{q(\mathbf{T})}{p(\mathbf{T}|\mathbf{X})} d\mathbf{T} = \\&= \mathcal{L}(q) + KL(q(\mathbf{T})||p(\mathbf{T}|\mathbf{X})) \geq \mathcal{L}(q).\end{aligned}$$

We would like to maximize lower bound  $\mathcal{L}(q)$ .

## Mean field approximation

### Independence assumption

$$q(\mathbf{T}) = \prod_{i=1}^k q_i(\mathbf{T}_i), \quad \mathbf{T} = [\mathbf{T}_1, \dots, \mathbf{T}_k], \quad \mathbf{T}_j = \{\mathbf{t}_{ij}\}_{i=1}^n, \quad \mathbf{t}_i = \{\mathbf{T}_{ij}\}_{j=1}^k.$$

Block coordinate optimization of ELBO for  $q_j(\mathbf{T}_j)$

$$\begin{aligned} \mathcal{L}(q) &= \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})}{q(\mathbf{T})} d\mathbf{T} = \int \left[ \prod_{i=1}^k q_i(\mathbf{T}_i) \right] \log \frac{p(\mathbf{X}, \mathbf{T})}{\left[ \prod_{i=1}^k q_i(\mathbf{T}_i) \right]} \prod_{i=1}^k d\mathbf{T}_i = \\ &= \int \left[ \prod_{i=1}^k q_i \right] \log p(\mathbf{X}, \mathbf{T}) \prod_{i=1}^k d\mathbf{T}_i - \sum_{i=1}^k \int \left[ \prod_{j=1}^k q_j \right] \log q_i \prod_{j=1}^k d\mathbf{T}_j = \\ &= \int q_j \left[ \int \log p(\mathbf{X}, \mathbf{T}) \prod_{i \neq j} q_i d\mathbf{T}_i \right] d\mathbf{T}_j - \\ &\quad - \int q_j \log q_j d\mathbf{T}_j + \text{const}(q_j) \rightarrow \max_{q_j} \end{aligned}$$

## Mean field approximation

Block coordinate optimization of ELBO for  $q_j(\mathbf{T}_j)$

$$\begin{aligned}\mathcal{L}(q) &= \int q_j \left[ \int \log p(\mathbf{X}, \mathbf{T}) \prod_{i \neq j} q_i d\mathbf{T}_i \right] d\mathbf{T}_j - \int q_j \log q_j d\mathbf{T}_j + \text{const}(q_j) = \\ &= \int q_j \log \hat{p}(\mathbf{X}, \mathbf{T}_j) d\mathbf{T}_j - \int q_j \log q_j d\mathbf{T}_j + \text{const}(q_j) \rightarrow \max_{q_j}.\end{aligned}$$

Here we introduce

$$\log \hat{p}(\mathbf{X}, \mathbf{T}_j) = \int \log p(\mathbf{X}, \mathbf{T}) \prod_{i \neq j} q_i d\mathbf{T}_i = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}(q_j)$$

Final ELBO derivation for  $q_j(\mathbf{T}_j)$

$$\begin{aligned}\mathcal{L}(q) &= \int q_j(\mathbf{T}_j) \log \hat{p}(\mathbf{X}, \mathbf{T}_j) d\mathbf{T}_j - \int q_j(\mathbf{T}_j) \log q_j(\mathbf{T}_j) d\mathbf{T}_j + \text{const}(q_j) = \\ &\quad \int q_j(\mathbf{T}_j) \log \frac{\hat{p}(\mathbf{X}, \mathbf{T}_j)}{q_j(\mathbf{T}_j)} d\mathbf{T}_j + \text{const}(q_j) = \\ &= -KL(q_j(\mathbf{T}_j) || \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.\end{aligned}$$

## Mean field approximation

Independence assumption

$$q(\mathbf{T}) = \prod_{i=1}^k q_i(\mathbf{T}_i), \quad \mathbf{T} = [\mathbf{T}_1, \dots, \mathbf{T}_k], \quad \mathbf{T}_j = \{\mathbf{t}_{ij}\}_{i=1}^n.$$

ELBO

$$\mathcal{L}(q) = -KL(q_j(\mathbf{T}_j) || \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.$$

Solution

$$q_j(\mathbf{T}_j) = \text{const} \cdot \hat{p}(\mathbf{X}, \mathbf{T}_j)$$

$$\log \hat{p}(\mathbf{X}, \mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

# Mean field approximation

## ELBO

$$\mathcal{L}(q) = -KL(q_j(\mathbf{T}_j) || \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.$$

## Solution

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

Assumptions:

- ▶  $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2] = [\mathbf{Z}, \boldsymbol{\theta}]$ ,  $q(\mathbf{T}) = q(\mathbf{T}_1) \cdot q(\mathbf{T}_2) = q(\mathbf{Z}) \cdot q(\boldsymbol{\theta})$ .
- ▶ restrict a class of probability distributions for  $\boldsymbol{\theta}$  to Dirac delta functions:

$$q_2 = q(\mathbf{T}_2) = q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*).$$

Under the restrictions the exact solution for  $q_2$  is not reached (KL can be greater than 0).

## Mean field approximation

General solution

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

Solution for  $q_1 = q(\mathbf{Z})$

$$\begin{aligned}\log q(\mathbf{Z}) &= \int q(\boldsymbol{\theta}) \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const} = \\ &= \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const} = \\ &= \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*) + \text{const.}\end{aligned}$$

EM-algorithm (E-step)

$$q(\mathbf{Z}) = \arg \max_q \mathcal{L}(q, \boldsymbol{\theta}^*) = \arg \min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*).$$

## Mean field approximation

### ELBO

$$\mathcal{L}(q) = -KL(q_j(\mathbf{T}_j) || \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.$$

ELBO maximization w.r.t.  $q_2 = q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$

$$\begin{aligned}\mathcal{L}(q_1, q_2) &= -KL(q(\boldsymbol{\theta}) || \hat{p}(\mathbf{X}, \boldsymbol{\theta})) + \text{const}(\boldsymbol{\theta}^*) \\ &= \int q(\boldsymbol{\theta}) \log \frac{\hat{p}(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}^*) \\ &= \int q(\boldsymbol{\theta}) \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}^*) \\ &= \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}^*) \rightarrow \max_{\boldsymbol{\theta}^*}\end{aligned}$$

## Mean field approximation

ELBO maximization w.r.t.  $q_2 = q(\theta) = \delta(\theta - \theta^*)$

$$\begin{aligned}\mathcal{L}(q_1, q_2) &= \int \delta(\theta - \theta^*) \log \hat{p}(\mathbf{X}, \theta) d\theta + \text{const} = \log \hat{p}(\mathbf{X}, \theta^*) + \text{const} \\ &= \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const} = \mathbb{E}_{q_1} \log p(\mathbf{X}, \mathbf{Z}, \theta^*) + \text{const} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} | \theta^*) d\mathbf{Z} + \log p(\theta^*) + \text{const} \rightarrow \max_{\theta^*}\end{aligned}$$

EM-algorithm (M-step)

$$\begin{aligned}\mathcal{L}(q, \theta) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} | \theta)}{q(\mathbf{Z})} d\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} | \theta) d\mathbf{Z} + \text{const} \rightarrow \max_{\theta}\end{aligned}$$

# Mean field approximation

## Solution

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

## EM algorithm (special case)

- ▶ Initialize  $\theta^*$ ;
- ▶ E-step

$$q(\mathbf{Z}) = \arg \max_q \mathcal{L}(q, \theta^*) = \arg \min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \theta^*);$$

- ▶ M-step
$$\theta^* = \arg \max_{\theta} \mathcal{L}(q, \theta);$$
- ▶ Repeat E-step and M-step until convergence.

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
- 4. IWAE**
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

## Theorem

1.  $\log p(\mathbf{x}|\theta) \geq \mathcal{L}_K(q, \theta) \geq \mathcal{L}_M(q, \theta)$ , for  $K \geq M$ ;
2.  $\log p(\mathbf{x}|\theta) = \lim_{K \rightarrow \infty} \mathcal{L}_K(q, \theta)$  if  $\frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})}$  is bounded.

## Proof of 1.

$$\begin{aligned}
\mathcal{L}_K(q, \theta) &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \log \left( \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \theta)}{q(\mathbf{z}_k | \mathbf{x})} \right) = \\
&= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \log \mathbb{E}_{k_1, \dots, k_M} \left( \frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_{k_m} | \theta)}{q(\mathbf{z}_{k_m} | \mathbf{x})} \right) \geq \\
&\geq \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \mathbb{E}_{k_1, \dots, k_M} \log \left( \frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_{k_m} | \theta)}{q(\mathbf{z}_{k_m} | \mathbf{x})} \right) = \\
&= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_M} \log \left( \frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_m | \theta)}{q(\mathbf{z}_m | \mathbf{x})} \right) = \mathcal{L}_M(q, \theta)
\end{aligned}$$

$$\frac{a_1 + \dots + a_K}{K} = \mathbb{E}_{k_1, \dots, k_M} \frac{a_{k_1} + \dots + a_{k_M}}{M}, \quad k_1, \dots, k_M \sim U[1, K]$$

## Theorem

1.  $\log p(\mathbf{x}|\theta) \geq \mathcal{L}_K(q, \theta) \geq \mathcal{L}_M(q, \theta)$ , for  $K \geq M$ ;
2.  $\log p(\mathbf{x}|\theta) = \lim_{K \rightarrow \infty} \mathcal{L}_K(q, \theta)$  if  $\frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})}$  is bounded.

## Proof of 2.

Consider r.v.  $\xi_K = \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\theta)}{q(\mathbf{z}_k|\mathbf{x})}$ .

If summands are bounded, then (from the strong law of large numbers)

$$\xi_K \xrightarrow[K \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} = p(\mathbf{x}|\theta).$$

Hence  $\mathcal{L}_K(q, \theta) = \mathbb{E} \log \xi_K$  converges to  $\log p(\mathbf{x}|\theta)$  as  $K \rightarrow \infty$ .

# Importance Weighted Autoencoders (IWAE)

## Theorem

1.  $\log p(\mathbf{x}|\theta) \geq \mathcal{L}_K(q, \theta) \geq \mathcal{L}_M(q, \theta)$ , for  $K \geq M$ ;
2.  $\log p(\mathbf{x}|\theta) = \lim_{K \rightarrow \infty} \mathcal{L}_K(q, \theta)$  if  $\frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x}, \phi)}$  is bounded.

If  $K > 1$  the bound could be tighter.

$$\mathcal{L}(q, \theta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x}, \phi)};$$

$$\mathcal{L}_K(q, \theta) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \phi)} \log \left( \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\theta)}{q(\mathbf{z}_k|\mathbf{x}, \phi)} \right).$$

- ▶  $\mathcal{L}_1(q, \theta) = \mathcal{L}(q, \theta)$ ;
- ▶  $\mathcal{L}_\infty(q, \theta) = \log p(\mathbf{x}|\theta)$ .
- ▶ Which  $q^*(\mathbf{z}|\mathbf{x}, \phi)$  gives  $\mathcal{L}(q^*, \theta) = \log p(\mathbf{x}|\theta)$ ?
- ▶ Which  $q^*(\mathbf{z}|\mathbf{x}, \phi)$  gives  $\mathcal{L}(q^*, \theta) = \mathcal{L}_K(q, \theta)$ ?

# Importance Weighted Autoencoders (IWAE)

## Theorem

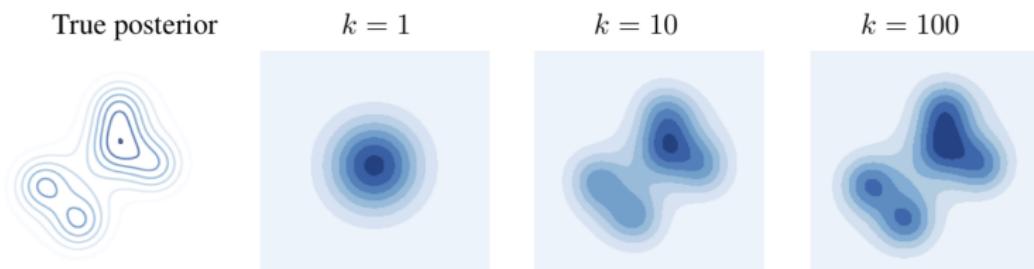
$\mathcal{L}(q^*, \theta) = \mathcal{L}_K(q, \theta)$  for the following variational distribution

$$q^*(\mathbf{z}|\mathbf{x}, \phi) = \mathbb{E}_{\mathbf{z}_2, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} q_{IW}(\mathbf{z}|\mathbf{x}, \mathbf{z}_{2:K}),$$

where

$$q_{IW}(\mathbf{z}|\mathbf{x}, \mathbf{z}_{2:K}) = \frac{\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}|\mathbf{x})}}{\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})}} q(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{\frac{1}{K} \left( \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + \sum_{k=2}^K \frac{p(\mathbf{x}, \mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})} \right)}.$$

## IWAE posterior



How to determine whether all VAE latent variables are informative?

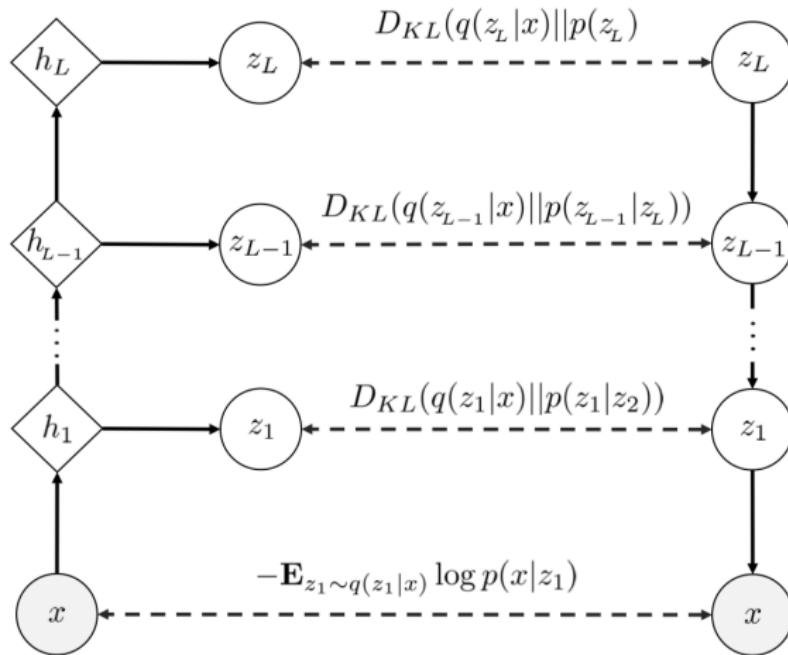
$$A_i = \text{cov}_{\mathbf{x}} (\mathbb{E}_{q(z_i|\mathbf{x})}[z_i]) > 0.01 \Leftrightarrow z_i \text{ is active}$$

| # stoch.<br>layers | k  | MNIST |                 |       |                 | OMNIGLOT |                 |        |                 |
|--------------------|----|-------|-----------------|-------|-----------------|----------|-----------------|--------|-----------------|
|                    |    | VAE   |                 | IWAE  |                 | VAE      |                 | IWAE   |                 |
|                    |    | NLL   | active<br>units | NLL   | active<br>units | NLL      | active<br>units | NLL    | active<br>units |
| 1                  | 1  | 86.76 | 19              | 86.76 | 19              | 108.11   | 28              | 108.11 | 28              |
|                    | 5  | 86.47 | 20              | 85.54 | 22              | 107.62   | 28              | 106.12 | 34              |
|                    | 50 | 86.35 | 20              | 84.78 | 25              | 107.80   | 28              | 104.67 | 41              |
| 2                  | 1  | 85.33 | 16+5            | 85.33 | 16+5            | 107.58   | 28+4            | 107.56 | 30+5            |
|                    | 5  | 85.01 | 17+5            | 83.89 | 21+5            | 106.31   | 30+5            | 104.79 | 38+6            |
|                    | 50 | 84.78 | 17+5            | 82.90 | 26+7            | 106.30   | 30+5            | 103.38 | 44+7            |

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
- 5. PixelVAE, Hierarchical VAE**
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

## Hierarchical VAE



## Hierarchical decomposition

$$p(\mathbf{z}_1, \dots, \mathbf{z}_L) = p(\mathbf{z}_L)p(\mathbf{z}_{L-1}|\mathbf{z}_L) \dots p(\mathbf{z}_1, \mathbf{z}_2);$$

$$q(\mathbf{z}_1, \dots, \mathbf{z}_L | \mathbf{x}) = q(\mathbf{z}_1 | \mathbf{x}) \dots q(\mathbf{z}_L | \mathbf{x}).$$

## ELBO

$$\begin{aligned} \mathcal{L}(q, \theta) &= \mathbb{E}_{q(\mathbf{z}_1 | \mathbf{x})} \log p(\mathbf{x} | \mathbf{z}_1, \theta) - KL(q(\mathbf{z}_1, \dots, \mathbf{z}_L | \mathbf{x}) || p(\mathbf{z}_1, \dots, \mathbf{z}_L)) \\ &= \mathbb{E}_{q(\mathbf{z}_1 | \mathbf{x})} \log p(\mathbf{x} | \mathbf{z}_1, \theta) - \int \prod_{j=1}^L q(\mathbf{z}_j | \mathbf{x}) \sum_{i=1}^L \log \frac{q(\mathbf{z}_i | \mathbf{x})}{p(\mathbf{z}_i | \mathbf{z}_{i+1})} d\mathbf{z}_1 \dots d\mathbf{z}_L \\ &= \mathbb{E}_{q(\mathbf{z}_1 | \mathbf{x})} \log p(\mathbf{x} | \mathbf{z}_1, \theta) - \sum_{i=1}^L \int \prod_{j=1}^L q(\mathbf{z}_j | \mathbf{x}) \log \frac{q(\mathbf{z}_i | \mathbf{x})}{p(\mathbf{z}_i | \mathbf{z}_{i+1})} d\mathbf{z}_1 \dots d\mathbf{z}_L \\ &= \mathbb{E}_{q(\mathbf{z}_1 | \mathbf{x})} \log p(\mathbf{x} | \mathbf{z}_1, \theta) - \sum_{i=1}^L \int q(\mathbf{z}_{i+1} | \mathbf{x}) q(\mathbf{z}_i | \mathbf{x}) \log \frac{q(\mathbf{z}_i | \mathbf{x})}{p(\mathbf{z}_i | \mathbf{z}_{i+1})} d\mathbf{z}_i d\mathbf{z}_{i+1} \\ &= \mathbb{E}_{q(\mathbf{z}_1 | \mathbf{x})} \log p(\mathbf{x} | \mathbf{z}_1, \theta) - \sum_{i=1}^L \mathbb{E}_{q(\mathbf{z}_{i+1} | \mathbf{x})} [KL(q(\mathbf{z}_i | \mathbf{x}) || p(\mathbf{z}_i | \mathbf{z}_{i+1}))] \end{aligned}$$

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
- 6. Posterior collapse**
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

## Posterior collapse: toy example

Let define latent variable model in the following way:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}$$

- ▶ prior distribution  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I})$ ;
- ▶ probabilistic model  $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{z}))$  (diagonal covariance);
- ▶ variational posterior  $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}))$  (diagonal covariance).

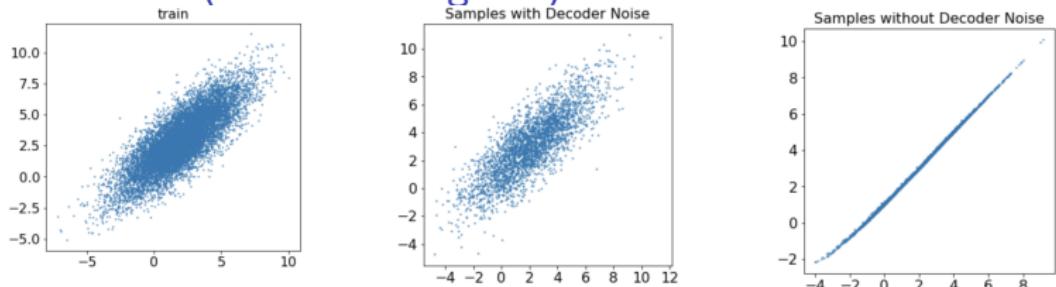
Let data distribution is  $\pi(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Possible cases:

- ▶ covariance matrix  $\boldsymbol{\Sigma}$  is diagonal (univariate case);
- ▶ covariance matrix  $\boldsymbol{\Sigma}$  is **not** diagonal (multivariate case).

What is the difference?

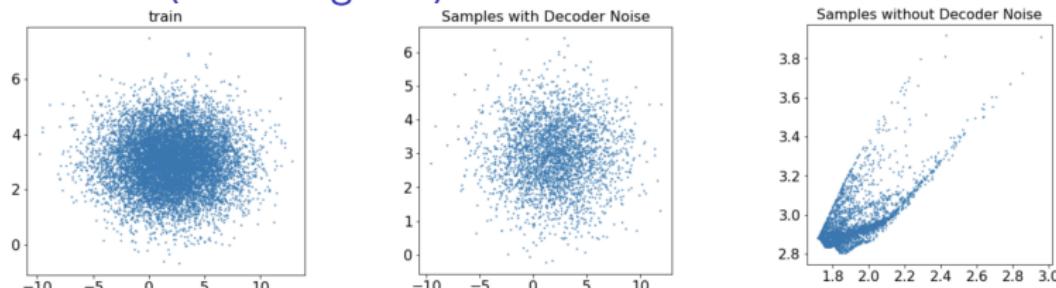
# Posterior collapse: toy example + VLAE

## Multivariate ( $\Sigma$ is non-diagonal)



The encoder uses latent variables to model data.

## Univariate ( $\Sigma$ is diagonal)



Latent variables are not used, since the decoder could model the data without the encoder.

# Variational Lossy AutoEncoder

Lossy code via explicit information placement

$$p(\mathbf{x}|\mathbf{z}, \theta) = \prod_{i=1}^m p(x_i|\mathbf{z}, \mathbf{x}_{\text{WindowAround}(i)}, \theta).$$

- ▶  $\text{WindowAround}(i)$  restricts the receptive field (it forbids to represent arbitrarily complex distribution over  $\mathbf{x}$  without dependence on  $\mathbf{z}$ ).
- ▶ Local statistics of 2D images (texture) will be modeled by a small local window.
- ▶ Global structural information (shapes) is long-range dependency that can only be communicated through latent code  $\mathbf{z}$ .

# Variational Lossy AutoEncoder

- ▶ Can VLAE learn lossy codes that encode global statistics?
- ▶ Does using AF priors improves upon using IAF posteriors as predicted by theory?
- ▶ Does using autoregressive decoding distributions improve density estimation performance?

## CIFAR10

### MNIST

| Model                                       | NLL Test     |
|---|--------------|
| Normalizing flows (Rezende & Mohamed, 2015) | 85.10        |
| DRAW (Gregor et al., 2015)                  | < 80.97      |
| Discrete VAE (Rollefson, 2016)              | 81.01        |
| PixelRNN (van den Oord et al., 2016a)       | 79.20        |
| IAF VAE (Kingma et al., 2016)               | 79.88        |
| AF VAE                                      | 79.30        |
| VLAE  | <b>79.03</b> |

| Method  | bits/dim $\leq$ |
|---|-----------------|
| <i>Results with tractable likelihood models:</i>                  |                 |
| Uniform distribution [1]  | 8.00            |
| Multivariate Gaussian [1]   | 4.70            |
| NICE [2]  | 4.48            |
| Deep GMMS [3]   | 4.00            |
| Real NVP [4]  | 3.49            |
| PixelCNN [1]  | 3.14            |
| Gated PixelCNN [5]  | 3.03            |
| PixelRNN [1]  | 3.00            |
| PixelCNN++ [6]  | <b>2.92</b>     |
| <i>Results with variationally trained latent-variable models:</i> |                 |
| Deep Diffusion [7]  | 5.40            |
| Convolutional DRAW [8]  | 3.58            |
| ResNet VAE with IAF [9]   | 3.11            |
| ResNet VLAE   | 3.04            |
| DenseNet VLAE   | <b>2.95</b>     |

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
- 7. Flows intuition**
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

## Flows intuition

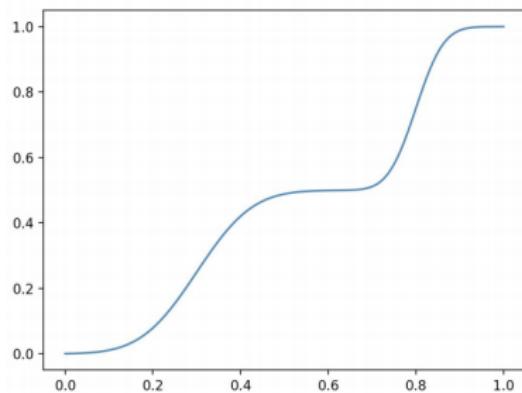
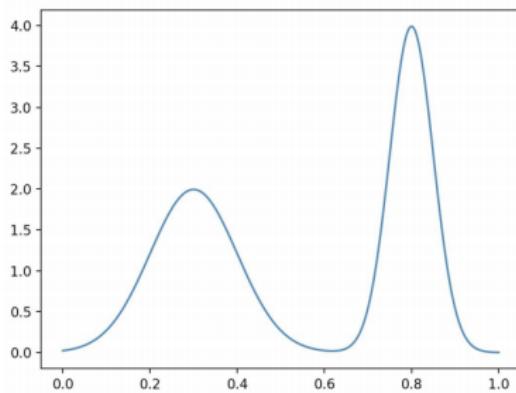
Let  $\xi$  be a random variable with density  $p(\xi)$ . Then

$$\eta = F(\xi) = \int_{-\infty}^{\xi} p(t)dt \sim U[0, 1].$$

$$P(\eta < y) = P(F(\xi) < y) = P(\xi < F^{-1}(y)) = F(F^{-1}(y)) = y$$

Hence

$$\eta \sim U[0, 1]; \quad \xi = F^{-1}(\eta) \quad \Rightarrow \quad \xi \sim p(\xi).$$



## Flows intuition

- ▶ Let  $z \sim p(z)$  is a random variable with base distribution  $p(z) = U[0, 1]$ .
- ▶ Let  $x \sim p(x)$  is a random variable with complex distribution  $p(x)$  and cdf  $F(x)$ .
- ▶ Then noise variable  $z$  can be transformed to  $x$  using inverse cdf  $F^{-1}$  ( $x = F^{-1}(z)$ ).

How to transform random variable  $z$  which has a distribution different from uniform to  $x$ ?

- ▶ Let  $z \sim p(z)$  is a random variable with base distribution  $p(z)$  and cdf  $G(z)$ .
- ▶ Then  $z_0 = G(z)$  has base distribution  $p(z_0) = U[0, 1]$ .
- ▶ Let  $x \sim p(x)$  is a random variable with complex distribution  $p(x)$  and cdf  $F(x)$ .
- ▶ Then noise variable  $z$  can be transformed to  $x$  using cdf  $G$  and inverse cdf  $F^{-1}$  ( $x = F^{-1}(z_0) = F^{-1}(G(z))$ ).

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
- 8. Parallel WaveNet**
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

# MAF/IAF pros and cons

## MAF

- ▶ Sampling is slow.
- ▶ Likelihood evaluation is fast.

## IAF

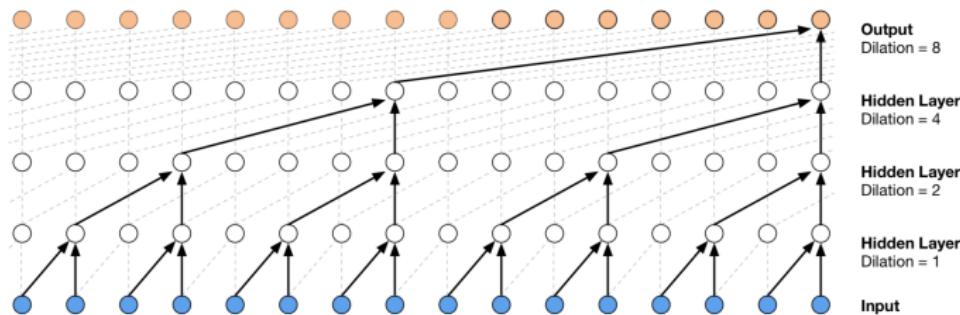
- ▶ Sampling is fast.
- ▶ Likelihood evaluation is slow.

How to take the best of both worlds?

## WaveNet

Autoregressive model with caused dilated convolutions for raw audio waveforms generation.

$$p(\mathbf{x}|\theta) = \prod_{t=1}^T p(x_t|\mathbf{x}_{1:t-1}, \theta).$$



## Parallel WaveNet

- ▶ 24kHz instead of 16kHz using increased dilated convolution filter size from 2 to 3.
- ▶ 16-bit signals with mixture of logistics instead of 8-bit signal with 256-way categorical distribution.

### Probability density distillation

1. Train usual WaveNet (MAF) via MLE (teacher network).
2. Train IAF WaveNet (student network), which attempts to match the probability of its own samples under the distribution learned by the teacher.

### Student objective

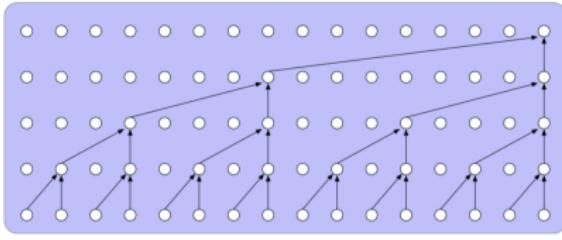
$$KL(p_s || p_t) = H(p_s, p_t) - H(p_s).$$

More than 1000x speed-up relative to original WaveNet!

# Parallel WaveNet

## WaveNet Teacher

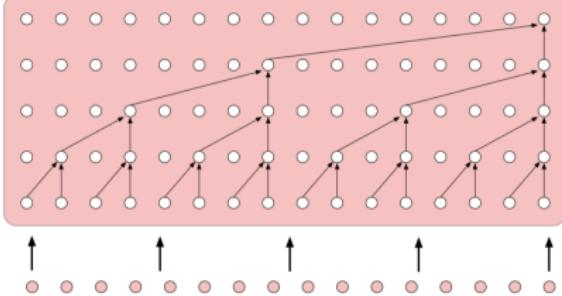
Linguistic features  $\dashrightarrow$



Teacher Output  
 $P(x_i | x_{<i})$

## WaveNet Student

Linguistic features  $\dashrightarrow$



Generated Samples  
 $x_i = g(z_i | z_{<i})$

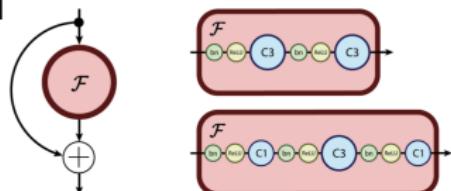
Student Output  
 $P(x_i | z_{<i})$

Input noise  
 $z_i$

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

- ▶ Modern neural networks are trained via backpropagation.
- ▶ Residual networks are state of the art in image classification.
- ▶ Backpropagation requires storing the network activations.



## Problem

Storing the activations imposes an increasing memory burden.

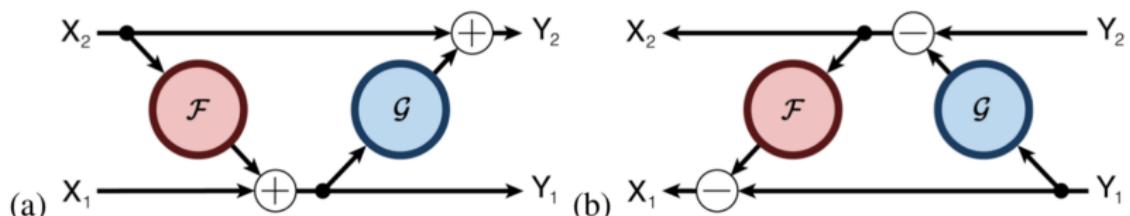
GPUs have limited memory capacity, leading to constraints often exceeded by state-of-the-art architectures (with thousand layers).

## NICE

$$\begin{cases} \mathbf{z}_1 = \mathbf{x}_1; \\ \mathbf{z}_2 = \mathbf{x}_2 + \mathcal{F}(\mathbf{x}_1, \theta); \end{cases} \Leftrightarrow \begin{cases} \mathbf{x}_1 = \mathbf{z}_1; \\ \mathbf{x}_2 = \mathbf{z}_2 - \mathcal{F}(\mathbf{z}_1, \theta). \end{cases}$$

## RevNet

$$\begin{cases} \mathbf{y}_1 = \mathbf{x}_1 + \mathcal{F}(\mathbf{x}_2, \theta); \\ \mathbf{y}_2 = \mathbf{x}_2 + \mathcal{G}(\mathbf{y}_1, \theta); \end{cases} \Leftrightarrow \begin{cases} \mathbf{x}_2 = \mathbf{y}_2 - \mathcal{F}(\mathbf{y}_1, \theta); \\ \mathbf{x}_1 = \mathbf{y}_1 - \mathcal{G}(\mathbf{x}_2, \theta). \end{cases}$$



| Architecture | CIFAR-10 [15] |              | CIFAR-100 [15] |               |
|--------------|---------------|--------------|----------------|---------------|
|              | ResNet        | RevNet       | ResNet         | RevNet        |
| 32 (38)      | <b>7.14%</b>  | 7.24%        | 29.95%         | <b>28.96%</b> |
| 110          | <b>5.74%</b>  | 5.76%        | 26.44%         | <b>25.40%</b> |
| 164          | 5.24%         | <b>5.17%</b> | <b>23.37%</b>  | 23.69%        |

- ▶ If the network contains non-reversible blocks (poolings, strides), activations for these blocks should be stored.
- ▶ To avoid storing activations in the modern frameworks, the backward pass should be manually redefined.

## Hypothesis

The success of deep convolutional networks is based on progressively discarding uninformative variability about the input with respect to the problem at hand.

- ▶ It is difficult to recover images from their hidden representations.
- ▶ Information bottleneck principle: an optimal representation must reduce the MI between an input and its representation to reduce uninformative variability + maximize the MI between the output and its representation to preserve each class from collapsing onto other classes.

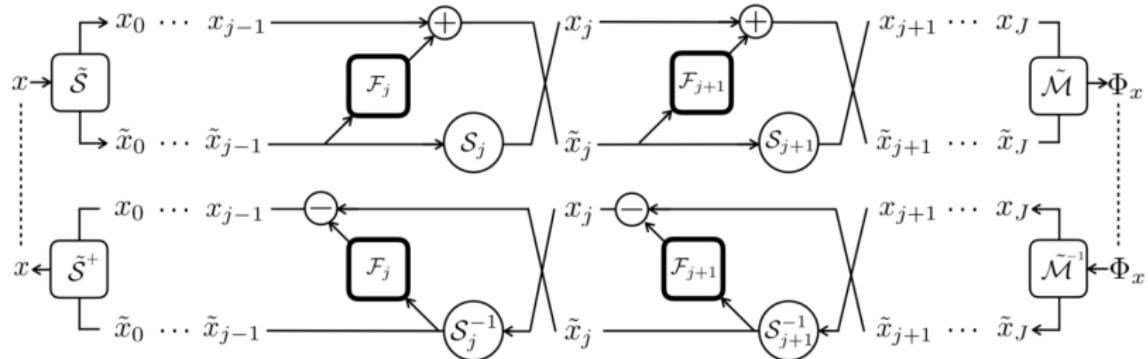
## Hypothesis

The success of deep convolutional networks is based on progressively discarding uninformative variability about the input with respect to the problem at hand.

## Idea

Build a cascade of homeomorphic layers (i-RevNet), a network that can be fully inverted up to the final projection onto the classes, i.e. no information is discarded.

# i-RevNet, 2018



| Architecture         | Injective | Bijective | Top-1 error | Parameters |
|----------------------|-----------|-----------|-------------|------------|
| ResNet               | -         | -         | 24.7        | 26M        |
| RevNet               | -         | -         | 25.2        | 28M        |
| <i>i</i> -RevNet (a) | yes       | -         | 24.7        | 181M       |
| <i>i</i> -RevNet (b) | yes       | yes       | 26.7        | 29M        |

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
- 10. ELBO surgery**
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

## ELBO interpretations

$$\log p(\mathbf{x}|\theta) = \mathcal{L}(q, \theta) + KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}|\mathbf{x}, \theta)).$$

$$\mathcal{L}(q, \theta) = \int q(\mathbf{z}|\mathbf{x}, \phi) \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x}, \phi)} d\mathbf{z}.$$

- ▶ Evidence minus posterior KL

$$\mathcal{L}(q, \theta) = \log p(\mathbf{x}|\theta) - KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}|\mathbf{x}, \theta)).$$

- ▶ Average negative energy plus entropy

$$\begin{aligned}\mathcal{L}(q, \theta) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} [\log p(\mathbf{x}, \mathbf{z}|\theta) - \log q(\mathbf{z}|\mathbf{x}, \phi)] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \log p(\mathbf{x}, \mathbf{z}|\theta) + \mathbb{H}[q(\mathbf{z}|\mathbf{x}, \phi)].\end{aligned}$$

- ▶ Average reconstruction minus KL to prior

$$\begin{aligned}\mathcal{L}(q, \theta) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} [\log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(\mathbf{z}) - \log q(\mathbf{z}|\mathbf{x}, \phi)] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z})).\end{aligned}$$

# ELBO surgery, 2016

$$\mathcal{L}(q, \theta) = \int q(\mathbf{Z}|\mathbf{X}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z}|\mathbf{X})} d\mathbf{Z}.$$

## ELBO interpretations

- ▶ Evidence minus posterior KL

$$\mathcal{L}(q, \theta) = \log p(\mathbf{X}|\theta) - KL(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z}|\mathbf{X}, \theta)).$$

- ▶ Average negative energy plus entropy

$$\mathcal{L}(q, \theta) = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})} p(\mathbf{X}, \mathbf{Z}|\theta) + \mathbb{H}[q(\mathbf{Z}|\mathbf{X})].$$

- ▶ Average term-by-term reconstruction minus KL to prior

$$\mathcal{L}(q, \theta) = \frac{1}{n} \sum_{i=1}^n [\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \theta) - KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i))].$$

# ELBO surgery, 2016

$$\mathcal{L}(q, \theta) = \frac{1}{n} \sum_{i=1}^n [\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \theta) - KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i))].$$

## Theorem

$$\frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) = KL(q(\mathbf{z})||p(\mathbf{z})) + \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}],$$

where  $i$  is treated as random variable:

$$q(i, \mathbf{z}) = q(i)q(\mathbf{z}|i); \quad p(i, \mathbf{z}) = p(i)p(\mathbf{z}); \quad q(i) = p(i) = \frac{1}{n}; \quad q(\mathbf{z}|i) = q(\mathbf{z}|\mathbf{x}_i).$$

$$q(\mathbf{z}) = \sum_{i=1}^n q(i, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i); \quad \mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}] = \mathbb{E}_{q(i,\mathbf{z})} \log \frac{q(i, \mathbf{z})}{q(i)q(\mathbf{z})}.$$

# ELBO surgery, 2016

## Theorem

$$\frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}_i | \mathbf{x}_i) || p(\mathbf{z}_i)) = KL(q(\mathbf{z}) || p(\mathbf{z})) + \mathbb{I}_{q(i, \mathbf{z})}[i, \mathbf{z}].$$

## Proof

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}_i | \mathbf{x}_i) || p(\mathbf{z}_i)) &= \sum_{i=1}^n \int q(i) q(\mathbf{z}|i) \log \frac{q(\mathbf{z}|i)}{p(\mathbf{z})} d\mathbf{z} = \\&= \sum_{i=1}^n \int q(i, \mathbf{z}) \log \frac{q(i, \mathbf{z})}{p(\mathbf{z})p(i)} d\mathbf{z} = \int \sum_{i=1}^n q(i, \mathbf{z}) \log \frac{q(\mathbf{z})q(i|\mathbf{z})}{p(\mathbf{z})p(i)} d\mathbf{z} = \\&= \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} + \int \sum_{i=1}^n q(i|\mathbf{z})q(\mathbf{z}) \log \frac{q(i|\mathbf{z})}{p(i)} d\mathbf{z} = \\&= KL(q(\mathbf{z}) || p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})} \mathbb{H}[q(i|\mathbf{z})] + \log n.\end{aligned}$$

# ELBO surgery, 2016

## Theorem

$$\frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}_i | \mathbf{x}_i) || p(\mathbf{z}_i)) = KL(q(\mathbf{z}) || p(\mathbf{z})) + \mathbb{I}_{q(i, \mathbf{z})}[i, \mathbf{z}].$$

## Proof (continued)

$$\frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}_i | \mathbf{x}_i) || p(\mathbf{z}_i)) = KL(q(\mathbf{z}) || p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})} \mathbb{H}[q(i|\mathbf{z})] + \log n$$

$$\begin{aligned}\mathbb{I}_{q(i, \mathbf{z})}[i, \mathbf{z}] &= \mathbb{E}_{q(i, \mathbf{z})} \log \frac{q(i, \mathbf{z})}{q(i)q(\mathbf{z})} = \mathbb{E}_{q(\mathbf{z})} \mathbb{E}_{q(i|\mathbf{z})} \log \frac{q(i|\mathbf{z})q(\mathbf{z})}{q(i)q(\mathbf{z})} = \\ &= \mathbb{E}_{q(\mathbf{z})} \mathbb{E}_{q(i|\mathbf{z})} \log \frac{q(i|\mathbf{z})}{q(i)} = -\mathbb{E}_{q(\mathbf{z})} \mathbb{H}[q(i|\mathbf{z})] + \log n.\end{aligned}$$

## Learnable VAE prior

### Optimal prior

$$KL(q_{\text{agg}}(\mathbf{z}) || p(\mathbf{z})) = 0 \Leftrightarrow p(\mathbf{z}) = q_{\text{agg}}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z} | \mathbf{x}_i).$$

### Mixture of Gaussians

$$p(\mathbf{z} | \boldsymbol{\lambda}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2), \quad \boldsymbol{\lambda} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^K.$$

### Variational Mixture of posteriors (VampPrior)

$$p(\mathbf{z} | \boldsymbol{\lambda}) = \frac{1}{K} \sum_{k=1}^K q(\mathbf{z} | \mathbf{u}_k),$$

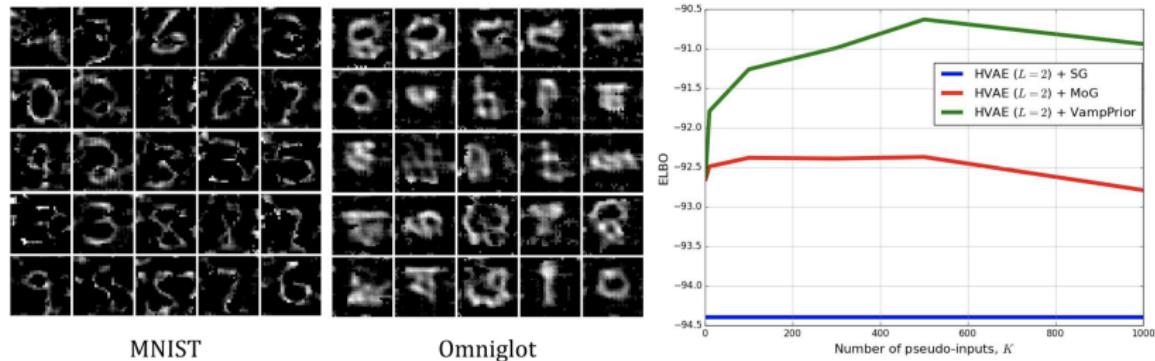
where  $\boldsymbol{\lambda} = \{\mathbf{u}_1, \dots, \mathbf{u}_K\}$  are trainable pseudo-inputs.

- ▶ Multimodal  $\Rightarrow$  prevents over-regularization;.
- ▶  $K \ll n \Rightarrow$  prevents from potential overfitting + less expensive to train.

# VampPrior

- ▶ Do we really need the multimodal prior?
- ▶ Is it beneficial to couple the prior with the variational posterior or the MoG prior is enough?

## Results



**Top row:** generated images by PixelHVAE + VampPrior for chosen pseudo-input in the left top corner.

**Bottom row:** pseudo-inputs for different datasets.

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
- 11. Disentanglement**
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

# InfoGAN

## GAN objective

$$\min_G \max_D V(G, D)$$

$$V(G, D) = \mathbb{E}_{\pi(x)} \log D(x) + \mathbb{E}_{p(z)} \log(1 - D(G(z)))$$

Latent vector  $\mathbf{z}$  is not imposed to be disentangled.

InfoGAN decomposes input vector:

- ▶  $\mathbf{z}$  – incompressible noise;
- ▶  $\mathbf{c}$  – structured latent code  $p(\mathbf{c}) = \prod_{j=1}^d p(c_j)$ .

## Information-theoretic regularization

$$\max I(\mathbf{c}, G(\mathbf{z}, \mathbf{c}))$$

Information in the latent code  $\mathbf{c}$  should not be lost in the  
generation process.

Chen X. et al. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, 2016

# InfoGAN

## Objective

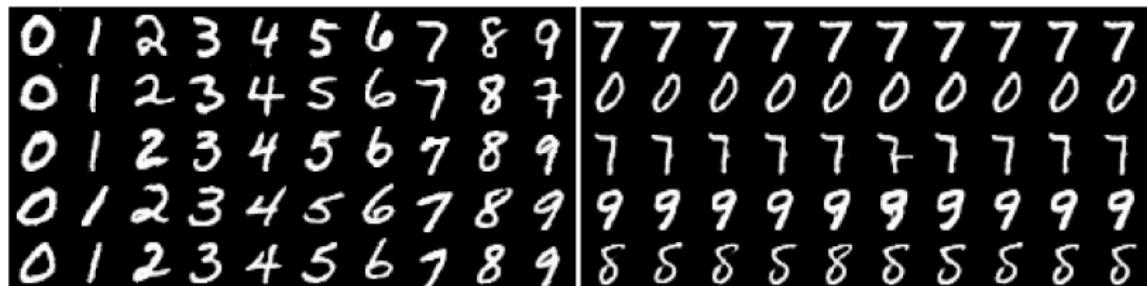
$$\min_G \max_D V(G, D) - \lambda I(\mathbf{c}, G(\mathbf{z}, \mathbf{c}))$$

## Variational Information Maximization

$$\begin{aligned} I(\mathbf{c}, G(\mathbf{z}, \mathbf{c})) &= H(\mathbf{c}) - H(\mathbf{c}|G(\mathbf{z}, \mathbf{c})) = \\ &= H(\mathbf{c}) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})} [\mathbb{E}_{\mathbf{c}' \sim p(\mathbf{c}|\mathbf{x})} \log p(\mathbf{c}'|\mathbf{x})] = \\ &= H(\mathbf{c}) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})} KL(p(\mathbf{c}'|\mathbf{x}) || q(\mathbf{z}'|\mathbf{x})) + \\ &\quad + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})} \mathbb{E}_{\mathbf{c}' \sim p(\mathbf{c}|\mathbf{x})} \log q(\mathbf{c}'|\mathbf{x}) \geq \\ &\geq H(\mathbf{c}) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})} \mathbb{E}_{\mathbf{c}' \sim p(\mathbf{c}|\mathbf{x})} \log q(\mathbf{c}'|\mathbf{x}) = \\ &\quad H(\mathbf{c}) + \mathbb{E}_{\mathbf{c} \sim p(\mathbf{c})} \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})} \log q(\mathbf{c}|\mathbf{x}) \end{aligned}$$

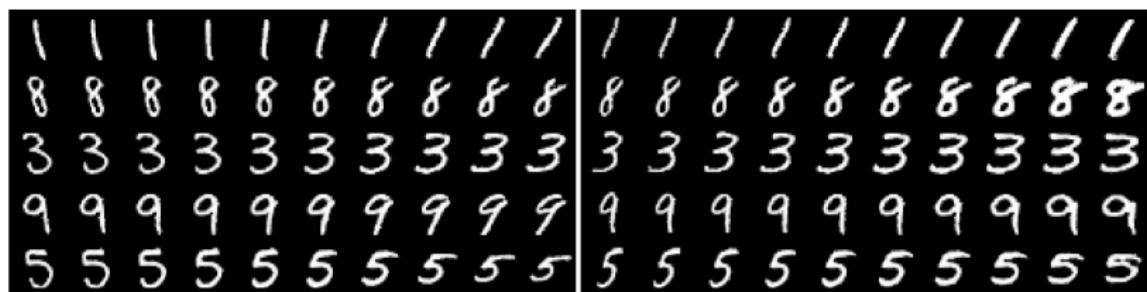
# InfoGAN

## Latent codes on MNIST



(a) Varying  $c_1$  on InfoGAN (Digit type)

(b) Varying  $c_1$  on regular GAN (No clear meaning)



(c) Varying  $c_2$  from  $-2$  to  $2$  on InfoGAN (Rotation)

(d) Varying  $c_3$  from  $-2$  to  $2$  on InfoGAN (Width)

# InfoGAN

## Latent codes on 3D Faces



(a) Azimuth (pose)

(b) Elevation



(c) Lighting

(d) Wide or Narrow

# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric**
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

# $\beta$ -VAE

## Disentangling metric

1. Generate two sets of objects

$$\mathbf{x}_{li} \sim \text{Sim}(\mathbf{v}_{li}, \mathbf{w}_{li}); \quad \mathbf{x}_{lj} \sim \text{Sim}(\mathbf{v}_{lj}, \mathbf{w}_{lj}); \quad y_{ij} \sim U[1, d].$$

$$\mathbf{v}_{li} \sim p(\mathbf{v}); \quad \mathbf{v}_{lj} \sim p(\mathbf{v}) ([v_{li}]_y = [v_{lj}]_y); \quad \mathbf{w}_{li}, \mathbf{w}_{lj} \sim p(\mathbf{w}).$$

2. Find representations

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu(\mathbf{x})|\sigma^2(\mathbf{x})); \quad \mathbf{z}_{li} = \mu(\mathbf{x}_{li}); \quad \mathbf{z}_{lj} = \mu(\mathbf{x}_{lj}).$$

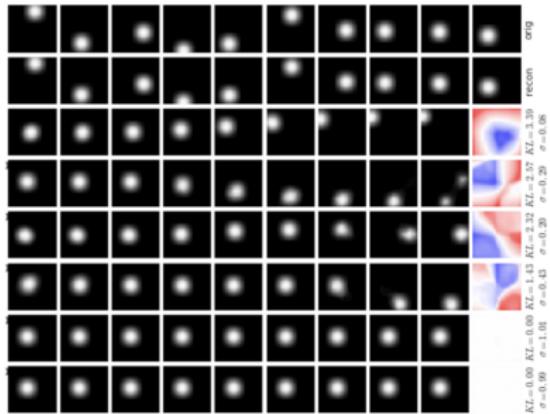
3. Use accuracy of classifier  $p(y|\mathbf{z}_{\text{diff}})$  with a low VC-dimension as metric of disentanglement

$$\mathbf{z}_{\text{diff}} = \frac{1}{L} \sum_{l=1}^L |\mathbf{z}_{li} - \mathbf{z}_{lj}|.$$

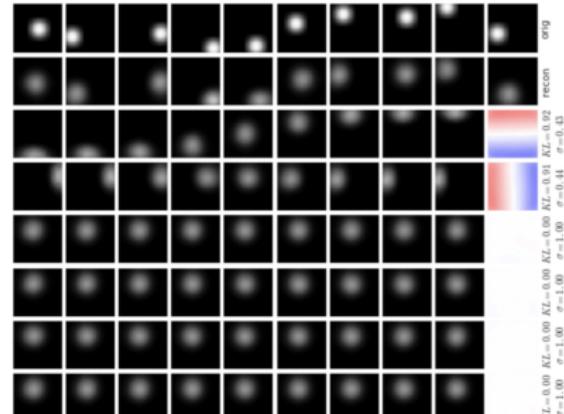
# $\beta$ -VAE

- ▶ **Top row:** original images.
- ▶ **Second row:** the corresponding reconstructions.
- ▶ **Remaining rows:** latent traversals ordered by KL divergence with the prior.
- ▶ **Heatmaps:** latent activations for each 2D position.

$\beta = 1$



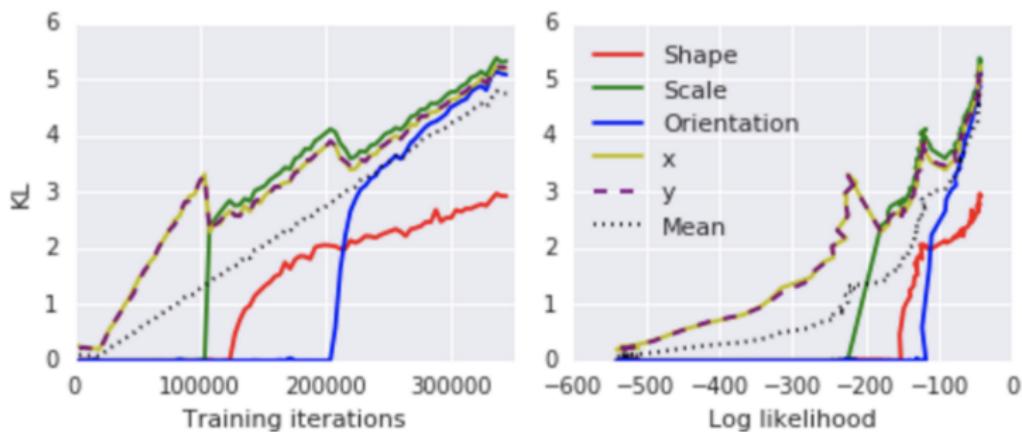
$\beta = 150$



# $\beta$ -VAE

## Controlled encoding capacity

$$\mathcal{L}(q, \theta, \beta) = \mathbb{E}_{q(z|x)} \log p(x|z, \theta) - [KL(q(z|x)||p(z)) - C].$$

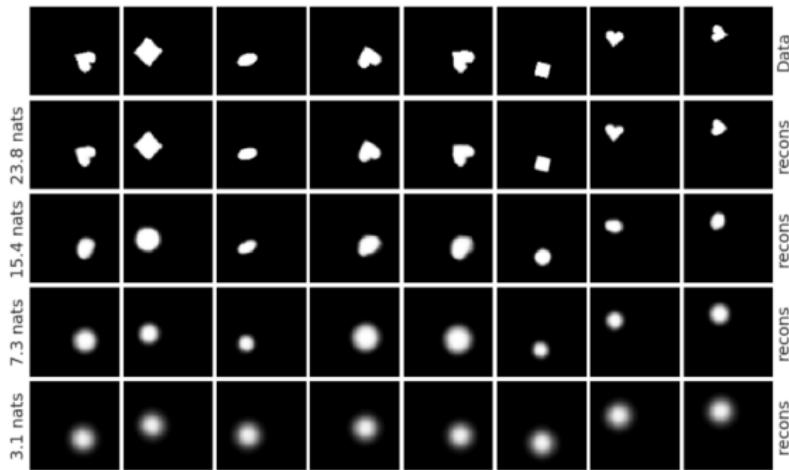


The early capacity is allocated to positional latents only, followed by a scale latent, then shape and orientation latents.

# $\beta$ -VAE

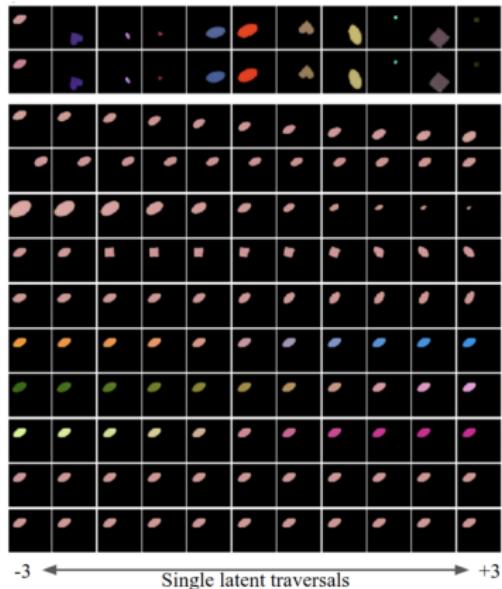
## Controlled encoding capacity

$$\mathcal{L}(q, \theta, \beta) = \mathbb{E}_{q(z|x)} \log p(x|z, \theta) - |KL(q(z|x)||p(z)) - C|.$$

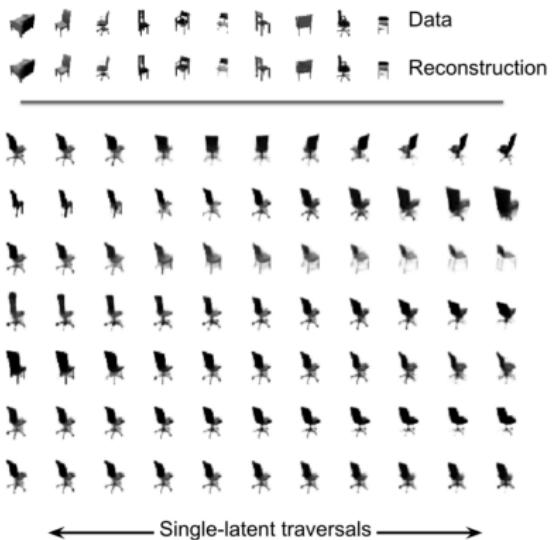


# $\beta$ -VAE

(a) Coloured dSprites



(b) 3D Chairs



# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

## FactorVAE

Disentangled aggregated variational posterior

$$q(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}) = \prod_{j=1}^d q(z_j)$$

Total correlation regularizer

$$\min KL(q(\mathbf{z}) || \prod_{j=1}^d q(z_j))$$

FactorVAE objective

$$\min_{\phi, \theta} \mathcal{L}(\phi, \theta) - \gamma \cdot KL(q(\mathbf{z}) || \prod_{j=1}^d q(z_j))$$

- ▶ The last term is intractable.
- ▶ FactorVAE uses density ratio trick for estimation.

## FactorVAE

Consider two distributions  $q_1(\mathbf{x})$ ,  $q_2(\mathbf{x})$  and probabilistic model

$$p(\mathbf{x}|y) = \begin{cases} q_1(\mathbf{x}), & \text{if } y = 1, \\ q_2(\mathbf{x}), & \text{if } y = 0, \end{cases} \quad y \sim \text{Bern}(0.5).$$

### Density ratio trick

$$\begin{aligned} \frac{q_1(\mathbf{x})}{q_2(\mathbf{x})} &= \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} = \frac{p(y=1|\mathbf{x})p(\mathbf{x})}{p(y=1)} \Big/ \frac{p(y=0|\mathbf{x})p(\mathbf{x})}{p(y=0)} = \\ &= \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \frac{p(y=1|\mathbf{x})}{1 - p(y=1|\mathbf{x})} = \frac{D(\mathbf{x})}{1 - D(\mathbf{x})} \end{aligned}$$

Here  $D(\mathbf{x})$  could be treated as a discriminator a model the output of which is a probability that  $\mathbf{x}$  is a sample from  $q_1(\mathbf{x})$  rather than from  $q_2(\mathbf{x})$ .

# FactorVAE

## FactorVAE objective

$$\min_{\theta, \phi} \text{ELBO}(\theta, \phi) - \gamma \cdot KL(q(\mathbf{z}) || \prod_{j=1}^d q(z_j))$$

## Total correlation regularizer

$$\begin{aligned} KL(q(\mathbf{z}) || \prod_{j=1}^d q(z_j)) &= KL(q(\mathbf{z}) || \bar{q}(\mathbf{z})) = \\ &= \mathbb{E}_{q(\mathbf{z})} \log \frac{q(\mathbf{z})}{\bar{q}(\mathbf{z})} \approx \mathbb{E}_{q(\mathbf{z})} \log \frac{D(\mathbf{z})}{1 - D(\mathbf{z})} \end{aligned}$$

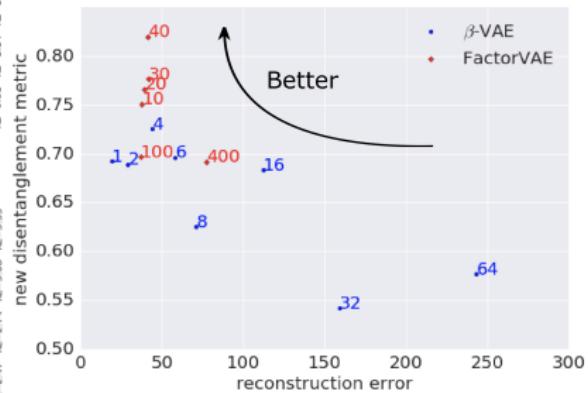
VAE and GAN are trained simultaneously.

# FactorVAE

$\beta$ -VAE ( $\beta = 8$ )



FactorVAE ( $\gamma = 10$ )



# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

# Challenging Disentanglement Assumptions

## Proof (1)

1. Consider the function  $g : \text{supp}(\mathbf{z}) \rightarrow [0, 1]^d$ :

$$g_i(\mathbf{u}) = P(z_i \leq u_i), \quad i = 1, \dots, d.$$

- ▶  $g$  is bijective (since  $p(\mathbf{z}) = \prod_{i=1}^d p(z_i)$ ).
  - ▶  $\frac{\partial g_i(\mathbf{u})}{\partial u_i} \neq 0$ , for all  $i$  and  $\frac{\partial g_i(\mathbf{u})}{\partial u_j} = 0$  for all  $i \neq j$ .
  - ▶  $g(\mathbf{z})$  is an independent  $d$ -dimensional uniform distribution.
2. Consider  $h : (0, 1]^d \rightarrow \mathbb{R}^d$

$$h_i(\mathbf{u}) = \psi^{-1}(u_i), \quad i = 1, \dots, d.$$

Here  $\psi$  denotes the CDF of a standard normal distribution.

- ▶  $h$  is bijective.
- ▶  $\frac{\partial h_i(\mathbf{u})}{\partial u_i} \neq 0$ , for all  $i$  and  $\frac{\partial h_i(\mathbf{u})}{\partial u_j} = 0$  for all  $i \neq j$ .
- ▶  $h(g(\mathbf{z}))$  is a  $d$ -dimensional standard normal distribution.

# Challenging Disentanglement Assumptions

## Proof (2)

Let  $\mathbf{A} \in \mathbb{R}^{d \times d}$  be an arbitrary orthogonal matrix with  $A_{ij} \neq 0$  for all  $i, j$ . The family of such matrices is infinite.

- ▶  $\mathbf{A}$  is orthogonal, it is invertible and thus defines a bijective linear operator.
- ▶  $\mathbf{A}h(g(\mathbf{z})) \in \mathbb{R}^d$  is hence an independent, multivariate standard normal distribution.
- ▶  $h^{-1}(\mathbf{A}h(g(\mathbf{z}))) \in \mathbb{R}^d$  is an independent  $d$ -dimensional uniform distribution.

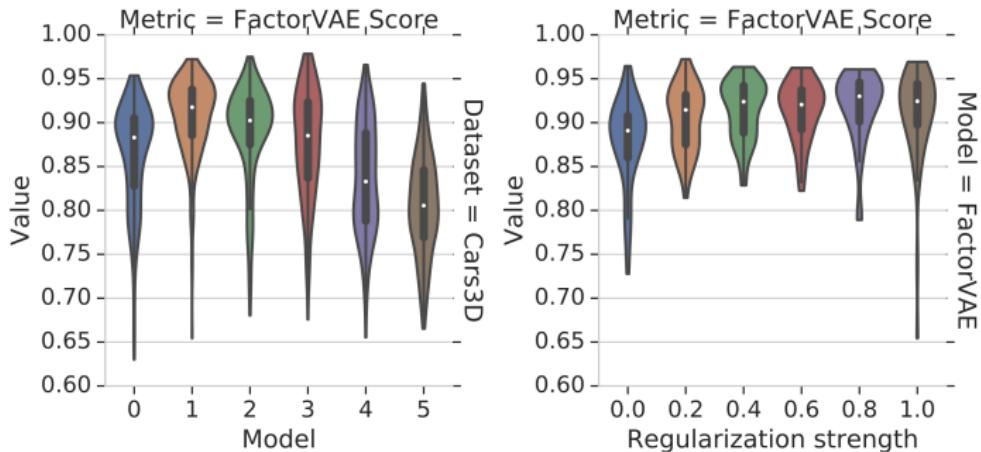
Define  $f : \text{supp}(\mathbf{z}) \rightarrow \text{supp}(\mathbf{z})$ :

$$f(\mathbf{u}) = g^{-1}(h^{-1}(\mathbf{A}h(g(\mathbf{z}))).$$

By definition  $f(\mathbf{z})$  has the same marginal distribution as  $\mathbf{z}$ :

$$P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u}) \text{ and } \frac{\partial f_i(\mathbf{z})}{\partial z_j} \neq 0.$$

# Challenging disentanglement assumptions



|                         | Dataset = Noisy-dSprites |     |     |     |     |     |
|-------------------------|--------------------------|-----|-----|-----|-----|-----|
| BetaVAE Score (A)       | 100                      | 80  | 44  | 41  | 46  | 37  |
| FactorVAE Score (B)     | 80                       | 100 | 49  | 52  | 25  | 38  |
| MIG (C)                 | 44                       | 49  | 100 | 76  | 6   | 42  |
| DCI Disentanglement (D) | 41                       | 52  | 76  | 100 | -8  | 38  |
| Modularity (E)          | 46                       | 25  | 6   | -8  | 100 | 13  |
| SAP (F)                 | 37                       | 38  | 42  | 38  | 13  | 100 |
|                         | (A)                      | (B) | (C) | (D) | (E) | (F) |

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
  - DCGAN
  - Improved techniques for training GANs
  - WGAN
  - StyleGAN
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

# Outline

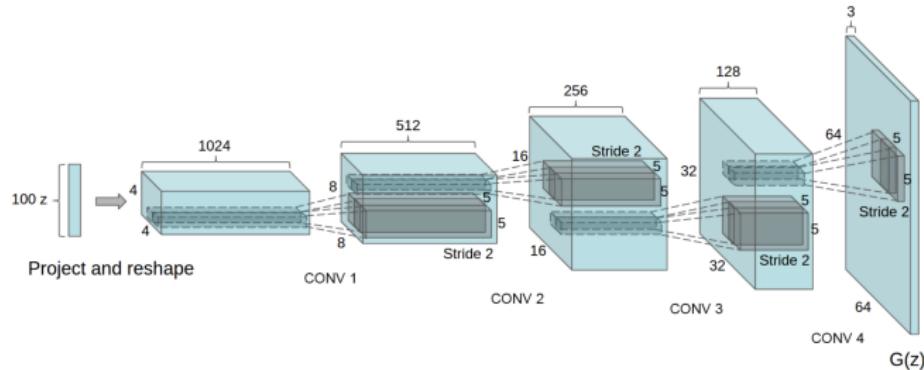
1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

## Vanilla GAN results



# Deep Convolutional GAN

## Architecture

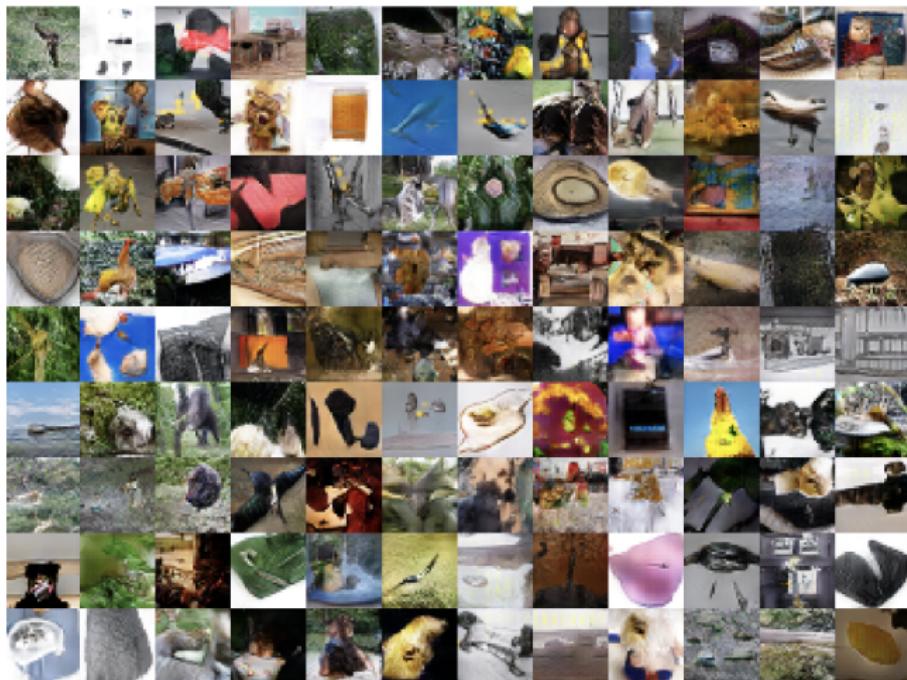


- ▶ Mean-pooling instead of max-pooling.
- ▶ Transposed convolutions in the generator for upsampling.
- ▶ Downsample with strided convolutions and average pooling.
- ▶ ReLU for generator, Leaky-ReLU (0.2) for discriminator.
- ▶ Output nonlinearity: tanh for Generator, sigmoid for discriminator.
- ▶ Batch Normalization used to prevent mode collapse (not applied at the output of  $G$  and input of  $D$ ).
- ▶ Adam: small LR = 2e-4; small momentum: 0.5, batch-size: 128.

*Radford A., Metz L., Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2015*

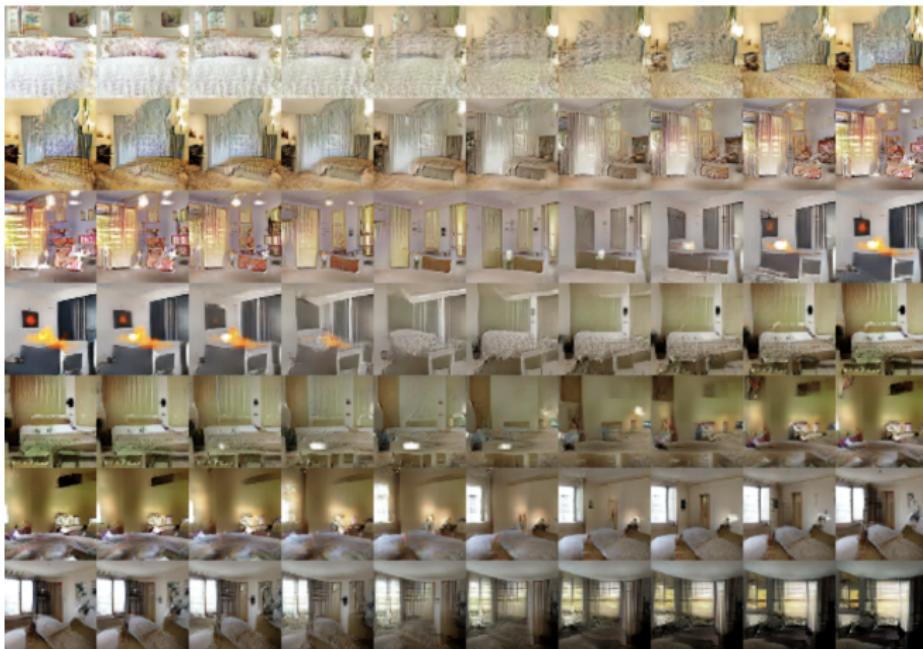
# Deep Convolutional GAN

## ImageNet samples



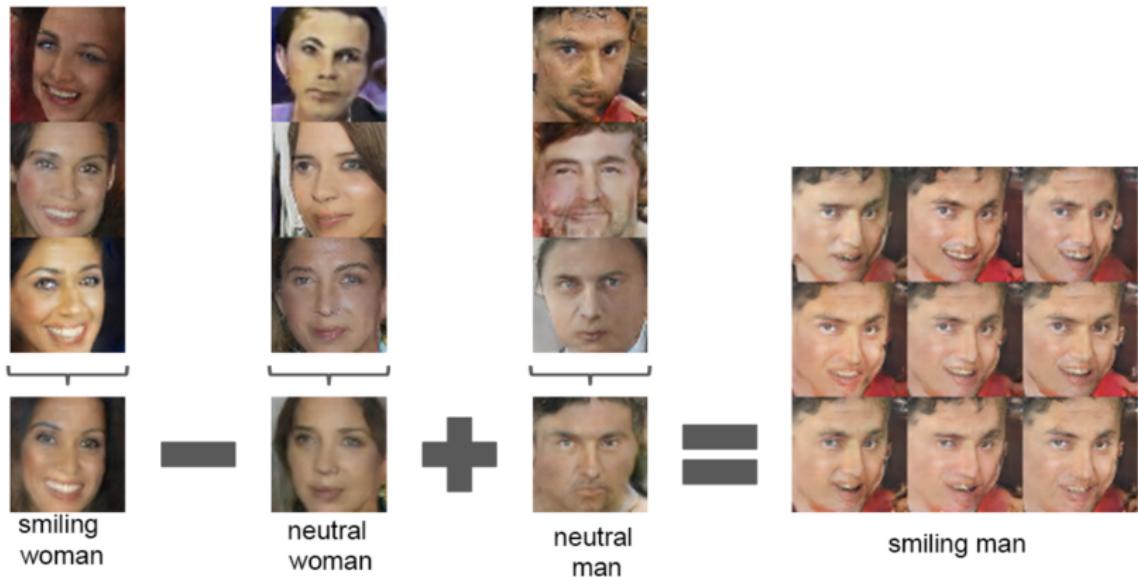
# Deep Convolutional GAN

## Smooth interpolations



# Deep Convolutional GAN

## Vector arithmetic



# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

# Improved techniques for training GANs

- ▶ Feature matching

$$\mathcal{L}_G = \|\mathbb{E}_{\pi(\mathbf{x})} \mathbf{d}(\mathbf{x}) - \mathbb{E}_{p(\mathbf{z})} \mathbf{d}(G(\mathbf{z}))\|_2^2$$

Here  $\mathbf{d}(\mathbf{x})$  – intermediate layer of discriminator. Matching the learned discriminator statistics instead of the output of the discriminator. Helps to avoid the vanishing gradients for sufficiently good discriminator.

- ▶ Historical averaging adds extra loss term for generator and discriminator losses

$$\|\boldsymbol{\theta} - \frac{1}{T} \sum_{t=1}^T \boldsymbol{\theta}_t\|_2^2.$$

Here  $\boldsymbol{\theta}_t$  – value of parameters at the previous step  $t$ . It allows to stabilize training procedure.

# Improved techniques for training GANs

- ▶ One-sided label smoothing. Instead of using one-hot labels in classification, use  $(1 - \alpha)$  for real data (the generated samples are not smoothed).

$$D^*(\mathbf{x}) = \frac{(1 - \alpha)\pi(\mathbf{x})}{\pi(\mathbf{x}) + p(\mathbf{x}|\theta)}$$

- ▶ Virtual batch normalization. BatchNorm makes samples within minibatch are highly correlated.



Use reference fixed batch to compute the normalization statistics. To avoid overfitting construct batch with the reference batch and the current sample.

# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

# Wasserstein 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  $t = 0, \dots, n_{\text{critic}}$  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:      $g_w \leftarrow \nabla_w [\frac{1}{m} \sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_w(g_\theta(z^{(i)}))]$ 
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
```

---

# Wasserstein GAN with Gradient Penalty

---

**Algorithm 1** WGAN with gradient penalty. We use default values of  $\lambda = 10$ ,  $n_{\text{critic}} = 5$ ,  $\alpha = 0.0001$ ,  $\beta_1 = 0$ ,  $\beta_2 = 0.9$ .

---

**Require:** The gradient penalty coefficient  $\lambda$ , the number of critic iterations per generator iteration  $n_{\text{critic}}$ , the batch size  $m$ , Adam hyperparameters  $\alpha, \beta_1, \beta_2$ .

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

```
1: while  $\theta$  has not converged do
2:   for  $t = 1, \dots, n_{\text{critic}}$  do
3:     for  $i = 1, \dots, m$  do
4:       Sample real data  $\mathbf{x} \sim \mathbb{P}_r$ , latent variable  $\mathbf{z} \sim p(\mathbf{z})$ , a random number  $\epsilon \sim U[0, 1]$ .
5:        $\tilde{\mathbf{x}} \leftarrow G_\theta(\mathbf{z})$ 
6:        $\hat{\mathbf{x}} \leftarrow \epsilon \mathbf{x} + (1 - \epsilon) \tilde{\mathbf{x}}$ 
7:        $L^{(i)} \leftarrow D_w(\tilde{\mathbf{x}}) - D_w(\mathbf{x}) + \lambda(\|\nabla_{\hat{\mathbf{x}}} D_w(\hat{\mathbf{x}})\|_2 - 1)^2$ 
8:     end for
9:      $w \leftarrow \text{Adam}(\nabla_w \frac{1}{m} \sum_{i=1}^m L^{(i)}, w, \alpha, \beta_1, \beta_2)$ 
10:   end for
11:   Sample a batch of latent variables  $\{\mathbf{z}^{(i)}\}_{i=1}^m \sim p(\mathbf{z})$ .
12:    $\theta \leftarrow \text{Adam}(\nabla_\theta \frac{1}{m} \sum_{i=1}^m -D_w(G_\theta(\mathbf{z})), \theta, \alpha, \beta_1, \beta_2)$ 
13: end while
```

---

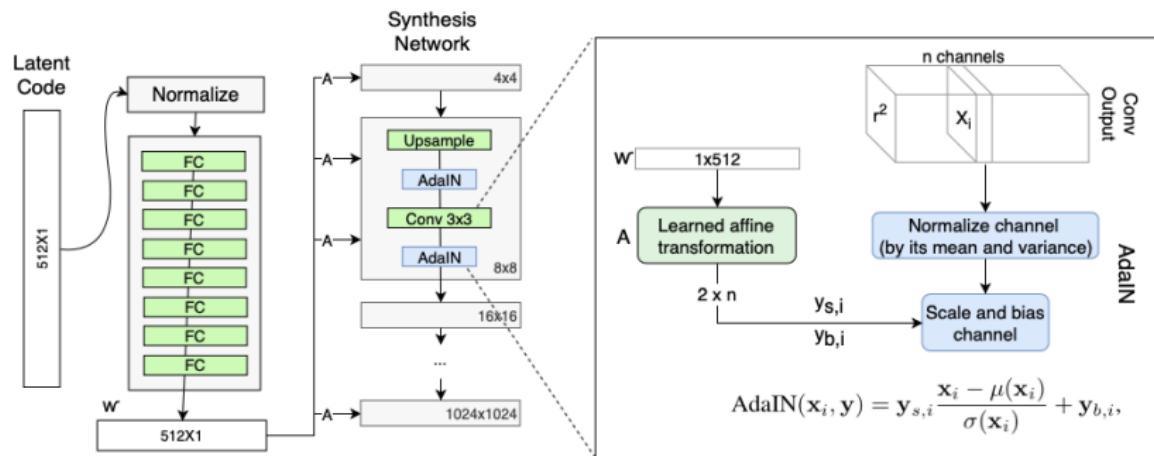
# Outline

1. Autoregressive models
  - Masked Autoencoder (MADE)
  - GatedPixelCNN
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
  - InfoGAN
  - $\beta$ -VAE, Disentanglement metric
  - FactorVAE
  - Challenging Disentanglement Assumptions
12. GANs
  - DCCGAN

# StyleGAN

## Step 2: Style modulation

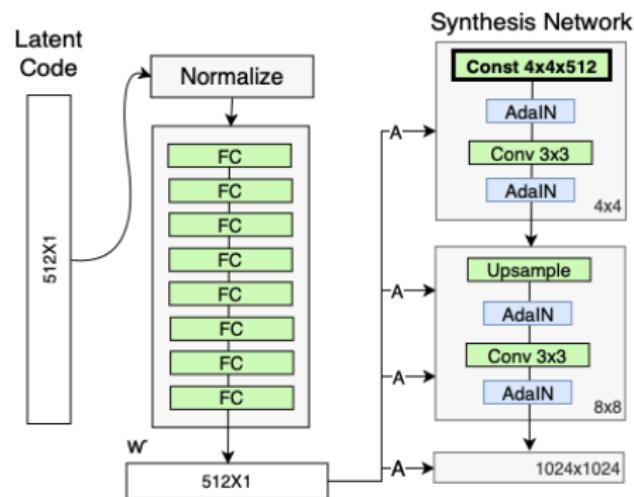
- ▶ Adaptive Instance Normalization transfers the  $\mathbf{w}$  vector to the synthesis Network.
- ▶ The module is added to each resolution to define the visual expression of the features.



# StyleGAN

## Step 3: Remove traditional input

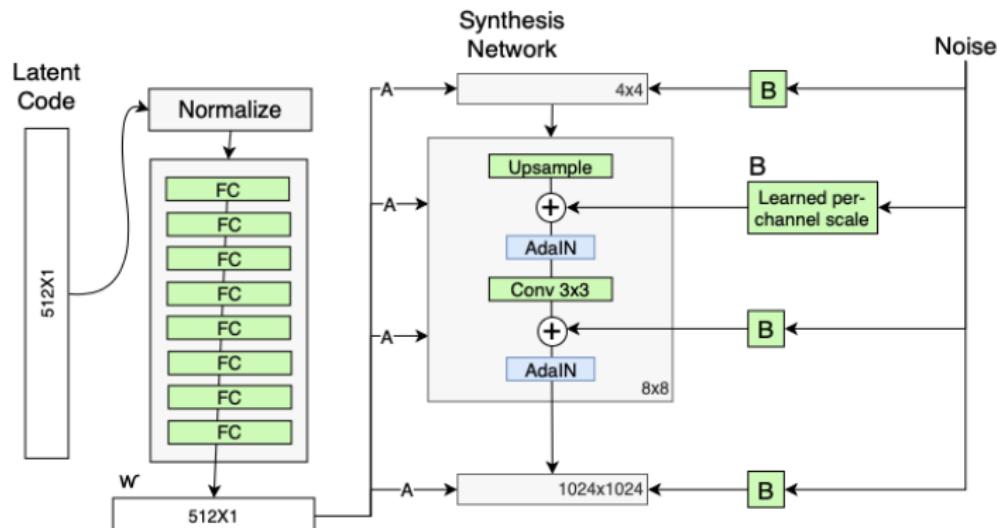
Mapping network provides stochasticity to different stages of the synthesis network. Input of the synthesis network is a trainable vector.



# StyleGAN

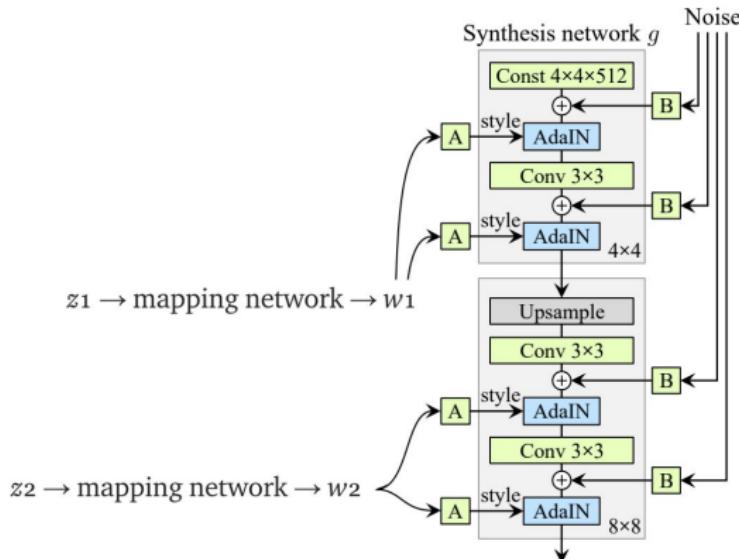
## Step 4: Stochastic variation

Inject random noise to add small aspects, such as freckles, exact placement of hairs, wrinkles, features which make the image more realistic and increase the variety of outputs.



# StyleGAN

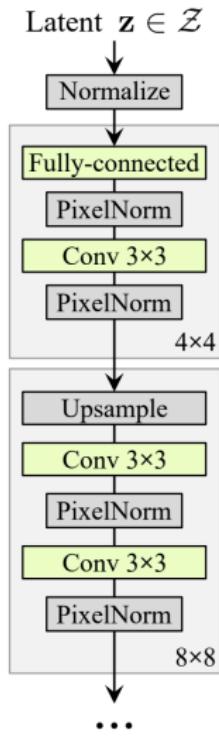
## Step 4: Style Mixing



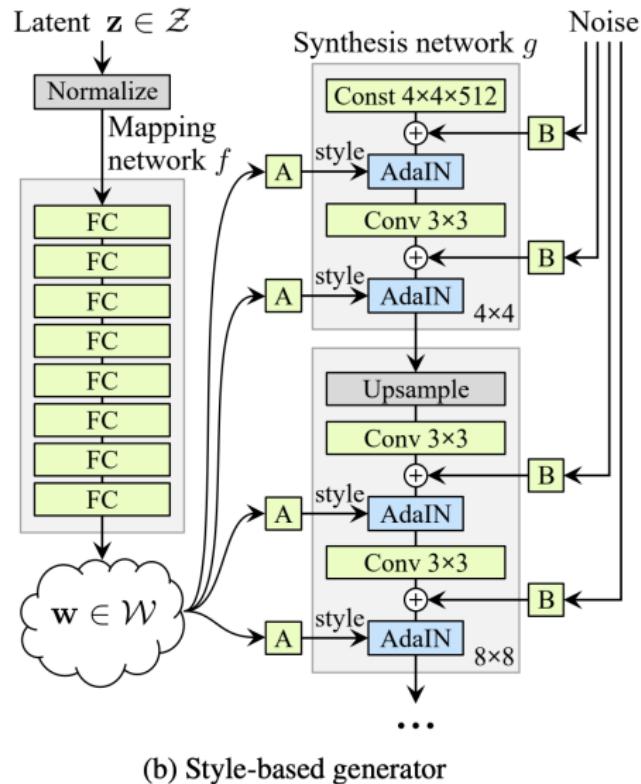
- ▶ Makes different levels of synthesis network to be independent.
- ▶ Allows to couple different styles.

Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks, 2018

# StyleGAN

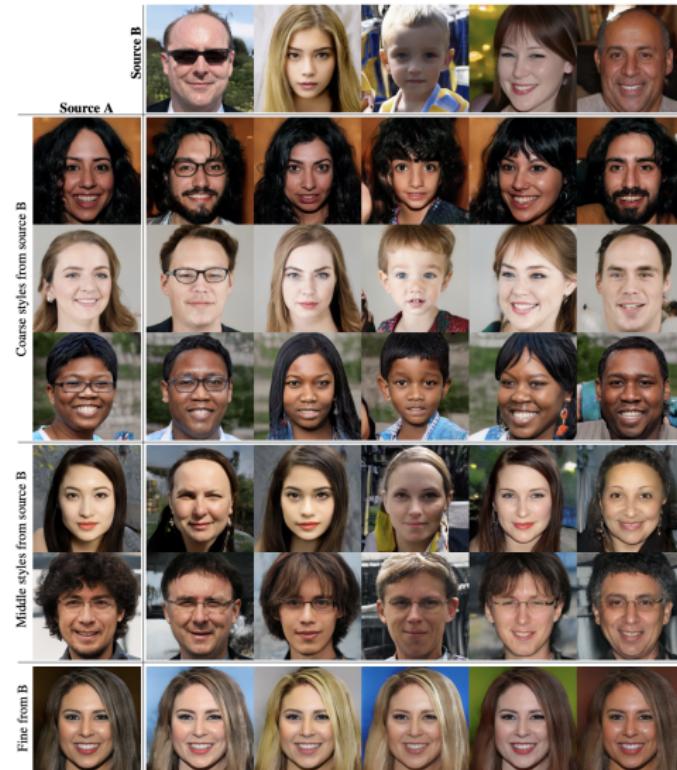


(a) Traditional



(b) Style-based generator

# StyleGAN



*Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks, 2018*

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
- 13. FFJORD**
14. Vector Quantized VAE-2
15. Feature Quantized GAN

# Continuous Normalizing Flows

Forward transform + log-density

$$\begin{bmatrix} \mathbf{x} \\ \log p(\mathbf{x}|\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \mathbf{z} \\ \log p(\mathbf{z}) \end{bmatrix} + \int_{t_0}^{t_1} \begin{bmatrix} f(\mathbf{z}(t), \boldsymbol{\theta}) \\ -\text{trace}\left(\frac{\partial f(\mathbf{z}(t), \boldsymbol{\theta})}{\partial \mathbf{z}(t)}\right) \end{bmatrix} dt.$$

- ▶ Discrete-in-time normalizing flows need invertible  $f$ . It costs  $O(d^3)$  to get determinant of Jacobian.
- ▶ Continuous-in-time flows require only smoothness of  $f$ . It costs  $O(d^2)$  to get trace of Jacobian.

It is possible to reduce cost from  $O(d^2)$  to  $O(d)$ !

Hutchinson's trace estimator

$$\text{trace}(A) = \mathbb{E}_{p(\epsilon)} \left[ \epsilon^T A \epsilon \right]; \quad \mathbb{E}[\epsilon] = 0; \quad \text{Cov}(\epsilon) = I.$$

FFJORD density estimation

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \mathbb{E}_{p(\epsilon)} \int_{t_0}^{t_1} \left[ \epsilon^T \frac{\partial f}{\partial \mathbf{z}} \epsilon \right] dt.$$

| Method                   |                                  | One-pass Sampling | Exact log-likelihood | Free-form Jacobian |
|--------------------------|----------------------------------|-------------------|----------------------|--------------------|
| Variational Autoencoders | Variational Autoencoders         | ✓                 | ✗                    | ✓                  |
|                          | Generative Adversarial Nets      | ✓                 | ✗                    | ✓                  |
|                          | Likelihood-based Autoregressive  | ✗                 | ✓                    | ✗                  |
| Change of Variables      | Normalizing Flows                | ✓                 | ✓                    | ✗                  |
|                          | Reverse-NF, MAF, TAN             | ✗                 | ✓                    | ✗                  |
|                          | NICE, Real NVP, Glow, Planar CNF | ✓                 | ✓                    | ✗                  |
|                          | <b>FFJORD</b>                    | ✓                 | ✓                    | ✓                  |

## Density estimation (forward KL)

|               | POWER        | GAS          | HEPMASS      | MINIBOONE    | BSDS300        | MNIST                             | CIFAR10      |
|---------------|--------------|--------------|--------------|--------------|----------------|-----------------------------------|--------------|
| Real NVP      | -0.17        | -8.33        | 18.71        | 13.55        | -153.28        | 1.06*                             | 3.49*        |
| Glow          | -0.17        | -8.15        | 18.92        | 11.35        | -155.07        | 1.05*                             | <b>3.35*</b> |
| <b>FFJORD</b> | <b>-0.46</b> | <b>-8.59</b> | <b>14.92</b> | <b>10.43</b> | <b>-157.40</b> | <b>0.99*</b> (1.05 <sup>†</sup> ) | 3.40*        |

## Flows for variational inference (reverse KL)

|               | MNIST                             | Omniglot                          | Frey Faces                       | Caltech Silhouettes                |
|---------------|-----------------------------------|-----------------------------------|----------------------------------|------------------------------------|
| IAF           | $84.20 \pm .17$                   | $102.41 \pm .04$                  | $4.47 \pm .05$                   | $111.58 \pm .38$                   |
| Sylvester     | $83.32 \pm .06$                   | $99.00 \pm .04$                   | $4.45 \pm .04$                   | $104.62 \pm .29$                   |
| <b>FFJORD</b> | <b><math>82.82 \pm .01</math></b> | <b><math>98.33 \pm .09</math></b> | <b><math>4.39 \pm .01</math></b> | <b><math>104.03 \pm .43</math></b> |

Grathwohl W. et al. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models, 2018

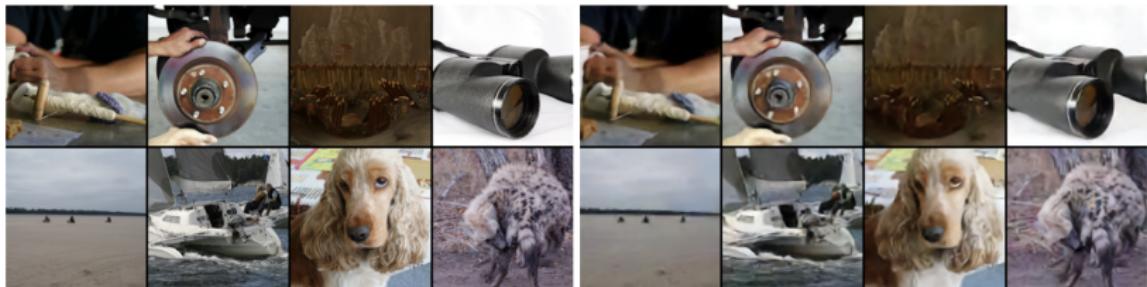
# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
- 14. Vector Quantized VAE-2**
15. Feature Quantized GAN

# Vector Quantized VAE

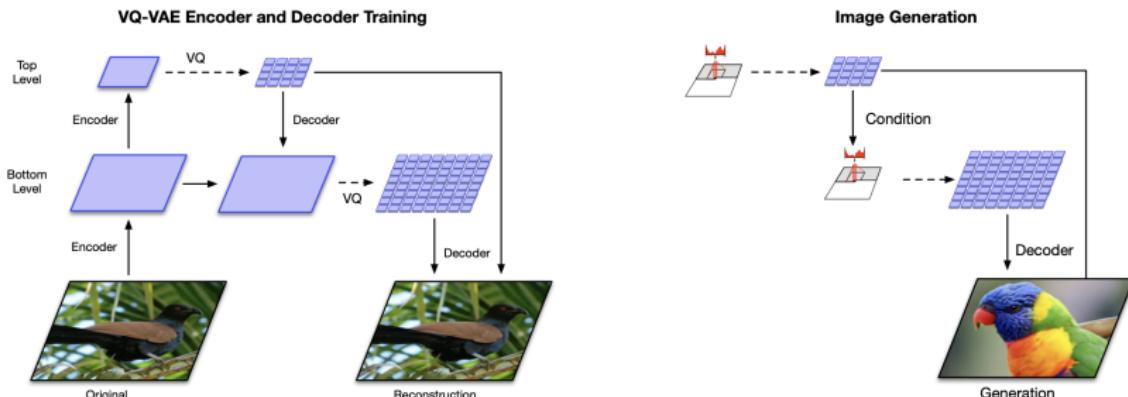
- ▶ The prior distribution over the discrete latents  $p(\hat{z})$  is a categorical distribution.
- ▶ It could be made autoregressive by depending on other  $\hat{z}$  in the feature map.
- ▶ While training the VQ-VAE, the prior is kept constant and uniform.
- ▶ After training, fit an autoregressive distribution (using PixelCNN) over  $\hat{z}$ .

## Samples



# Vector Quantized VAE-2

- ▶ Use multi-scale hierarchical model.
- ▶ Use autoregressive prior model in each scale of the hierarchy.
- ▶ Improve autoregressive prior (PixelSNAIL with self-attention in bottom layer, PixelCNN++ in bottom layer).
- ▶ Train the encoder and decoder at the first stage, train the priors at the second stage.



# Vector Quantized VAE-2

---

**Algorithm 1** VQ-VAE training (stage 1)

**Require:** Functions  $E_{top}$ ,  $E_{bottom}$ ,  $D$ ,  $\mathbf{x}$   
(batch of training images)

- 1:  $\mathbf{h}_{top} \leftarrow E_{top}(\mathbf{x})$   
    ▷ quantize with top codebook eq 1
- 2:  $\mathbf{e}_{top} \leftarrow Quantize(\mathbf{h}_{top})$
- 3:  $\mathbf{h}_{bottom} \leftarrow E_{bottom}(\mathbf{x}, \mathbf{e}_{top})$   
    ▷ quantize with bottom codebook eq 1
- 4:  $\mathbf{e}_{bottom} \leftarrow Quantize(\mathbf{h}_{bottom})$
- 5:  $\hat{\mathbf{x}} \leftarrow D(\mathbf{e}_{top}, \mathbf{e}_{bottom})$   
    ▷ Loss according to eq 2
- 6:  $\theta \leftarrow Update(\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}))$

---

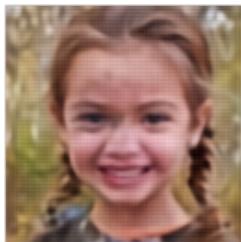
**Algorithm 2** Prior training (stage 2)

- 1:  $\mathbf{T}_{top}, \mathbf{T}_{bottom} \leftarrow \emptyset$                           ▷ training set
- 2: **for**  $\mathbf{x} \in$  training set **do**
- 3:      $\mathbf{e}_{top} \leftarrow Quantize(E_{top}(\mathbf{x}))$
- 4:      $\mathbf{e}_{bottom} \leftarrow Quantize(E_{bottom}(\mathbf{x}, \mathbf{e}_{top}))$
- 5:      $\mathbf{T}_{top} \leftarrow \mathbf{T}_{top} \cup \mathbf{e}_{top}$
- 6:      $\mathbf{T}_{bottom} \leftarrow \mathbf{T}_{bottom} \cup \mathbf{e}_{bottom}$
- 7: **end for**
- 8:  $p_{top} = TrainPixelCNN(\mathbf{T}_{top})$
- 9:  $p_{bottom} = TrainCondPixelCNN(\mathbf{T}_{bottom}, \mathbf{T}_{top})$

    ▷ Sampling procedure

- 10: **while** true **do**
- 11:      $\mathbf{e}_{top} \sim p_{top}$
- 12:      $\mathbf{e}_{bottom} \sim p_{bottom}(\mathbf{e}_{top})$
- 13:      $\mathbf{x} \leftarrow D(\mathbf{e}_{top}, \mathbf{e}_{bottom})$
- 14: **end while**

---

 $h_{top}$  $h_{top}, h_{middle}$  $h_{top}, h_{middle}, h_{bottom}$ 

Original

# Outline

1. Autoregressive models
2. ELBO gradient, Log derivative trick
3. Mean field approximation
4. IWAE
5. PixelVAE, Hierarchical VAE
6. Posterior collapse
7. Flows intuition
8. Parallel WaveNet
9. RevNet, i-RevNet
10. ELBO surgery
11. Disentanglement
12. GANs
13. FFJORD
14. Vector Quantized VAE-2
15. Feature Quantized GAN

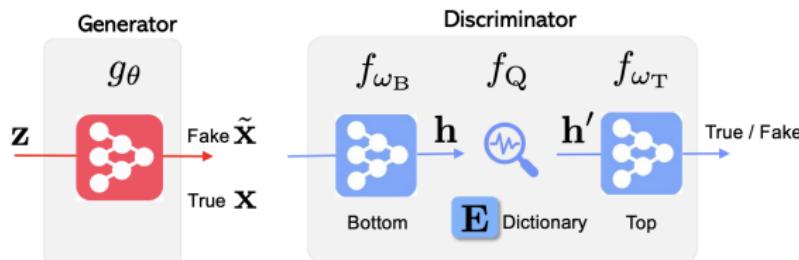
# Feature Quantized GAN

- ▶ GAN tries to find Nash equilibrium, minibatch training is unstable. GAN relies heavily on the minibatch statistics.
- ▶ Lots of feature matching strategies were proposed to stabilize the training.

## Feature quantized GAN discriminator

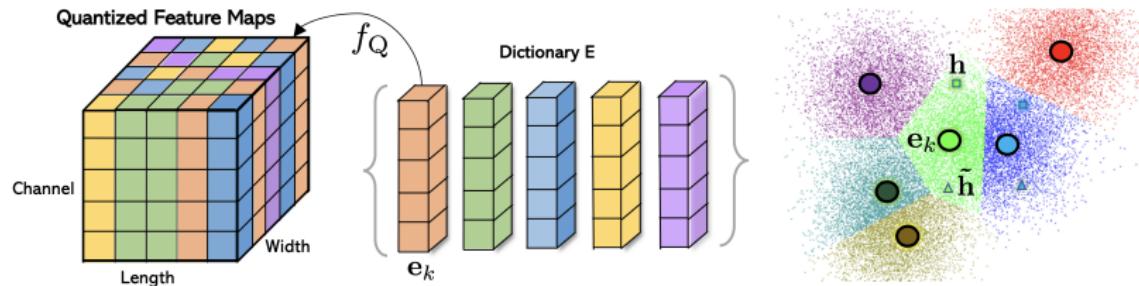
$$D(\mathbf{x}) = f_{\mathbf{w}_T} \circ f_{\mathbf{w}_B}(\mathbf{x}) \quad \Rightarrow \quad D(\mathbf{x}) = f_{\mathbf{w}_T} \circ f_Q \circ f_{\mathbf{w}_B}(\mathbf{x}).$$

Here  $f_Q$  is a vector quantization operation.



# Feature Quantized GAN

## Quantization procedure



## Quantized features



# Feature Quantized GAN

## ImageNet-1000

| Models | 64 × 64        |                           | 128 × 128                       |  |
|--------|----------------|---------------------------|---------------------------------|--|
|        | FID* ↓ / IS* ↑ |                           | FID* ↓ / IS* ↑                  |  |
| Half   | TAC-GAN        | -                         | 23.75 / 28.86±0.29 <sup>‡</sup> |  |
|        | BigGAN         | 12.75 / 21.84±0.34        | 22.77 / 38.05±0.79 <sup>‡</sup> |  |
| 256K   | FQ-BigGAN      | <b>12.62 / 21.99±0.32</b> | <b>19.11 / 41.92±1.15</b>       |  |
|        | BigGAN         | 10.55 / 25.43±0.15        | 14.88 / 63.03±1.42 <sup>†</sup> |  |
|        | FQ-BigGAN      | <b>9.67 / 25.96±0.24</b>  | <b>13.77</b> / 54.36±1.07       |  |

## FFHQ

| Resolution  | 32 <sup>2</sup> | 64 <sup>2</sup> | 128 <sup>2</sup> | 1024 <sup>2</sup> |
|-------------|-----------------|-----------------|------------------|-------------------|
| StyleGAN    | 3.28            | 4.82            | 6.33             | 5.24              |
| FQ-StyleGAN | <b>3.01</b>     | <b>4.36</b>     | <b>5.98</b>      | <b>4.89</b>       |

## Per class metrics for ImageNet

