
ETH zürich

From Variational to Deterministic Autoencoders

or the joys of density estimation in latent spaces

Antonio Vergari
University of California, Los Angeles

@tetraduzione

Joint work with: **Partha Ghosh, Mehdi S.M. Sajjadi, Bernhard Schölkopf, Michael Black**

26th August 2020 - **UCL - AI Center** Seminars

Why?

$$\{\mathbf{x}_i\}_{i=1}^{N} \sim p^*(\mathbf{X})$$

Why?

learning time

$$\{\mathbf{x}_i\}_{i=1}^{N} \sim p^*(\mathbf{X})$$

$$p_\theta(\mathbf{X})$$

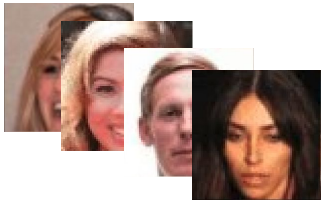
Why?

learning time

$$\{\mathbf{x}_i\}_{i=1}^{N} \sim p^*(\mathbf{X})$$

$$p_\theta(\mathbf{X})$$

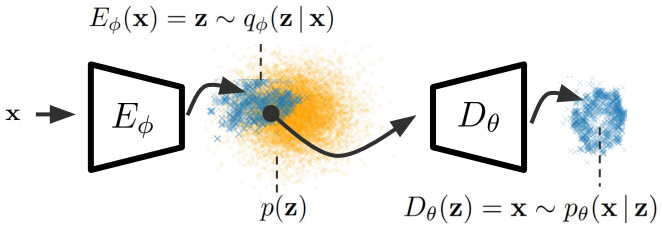
inference time



$$\mathbf{x}^{\text{new}} \sim p_\theta(\mathbf{X})$$

the generative modeling paradigm

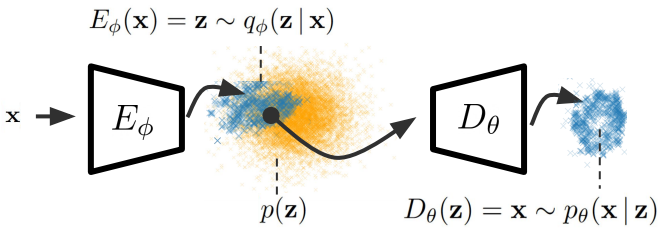
Variational Autoencoders
(VAEs)



- ⇒ ***Generative modeling*** *[Van Den Oord2017, Tolstikhin2019, Razavi2019,...]*
- ⇒ Density Estimation *[Kingma2014,Rezende2014,Burda2015,...]*
- ⇒ Disentanglement *[Higgings2016, ...]*

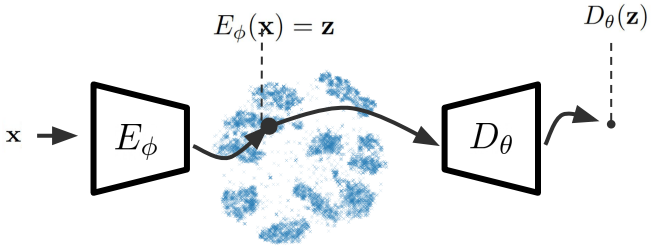
Kingma, Diederik P., and Max Welling. "***Auto-encoding variational bayes.***" ICLR 2014

Variational Autoencoders
(VAEs)



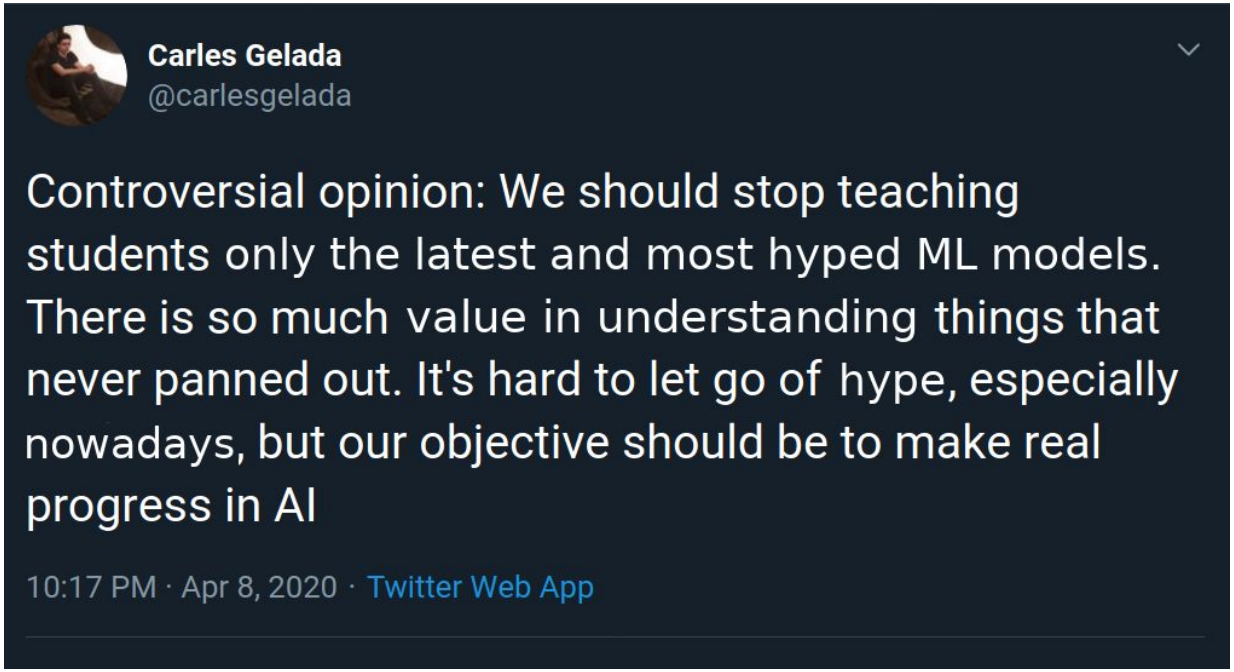
$\lesssim$

Regularized Autoencoders
(RAEs)



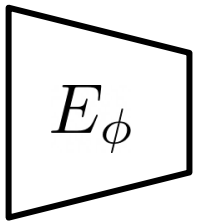
⇒ ***Generative modeling*** *[Van Den Oord2017, Tolstikhin2019, Razavi2019,...]*
⇒ Density Estimation *[Kingma2014,Rezende2014,Burda2015,...]*
⇒ Disentanglement *[Higgings2016, ...]*

a simpler alternative for generative modeling


Carles Gelada
@carlesgelada
Controversial opinion: We should stop teaching students only the latest and most hyped ML models. There is so much value in understanding things that never panned out. It's hard to let go of hype, especially nowadays, but our objective should be to make real progress in AI
10:17 PM · Apr 8, 2020 · Twitter Web App

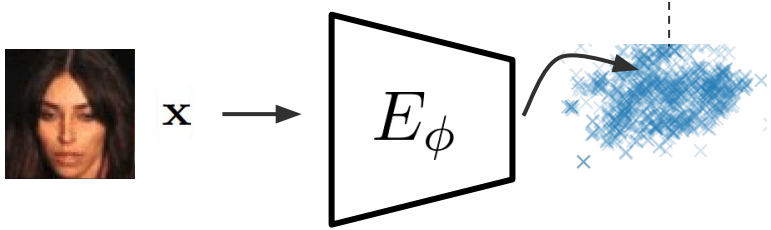
! disclaimer !

Variational Autoencoders (VAEs)

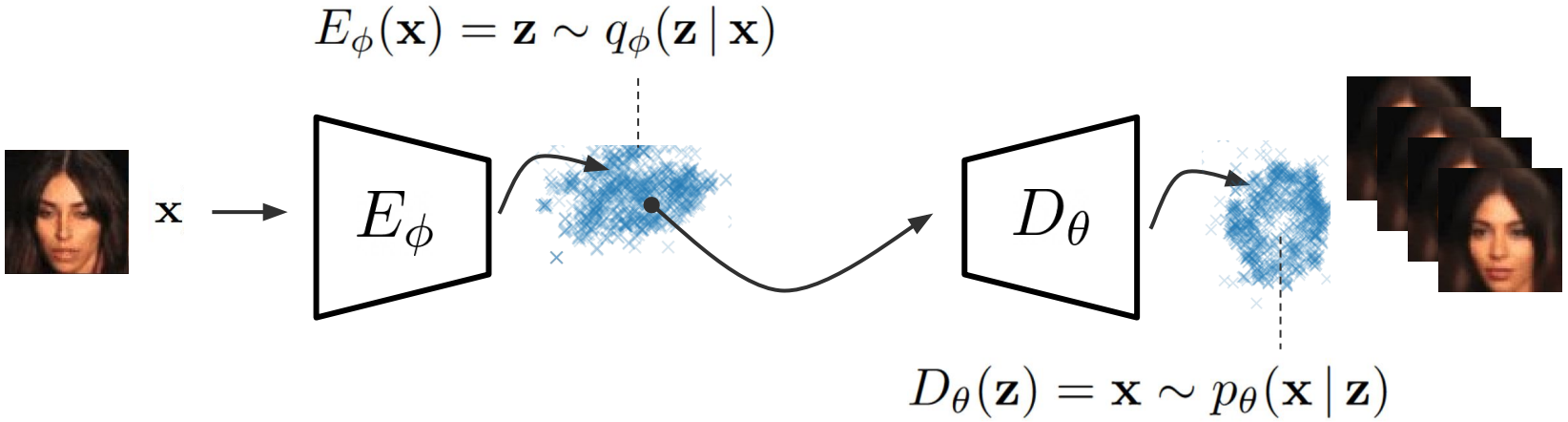


Variational Autoencoders (VAEs)

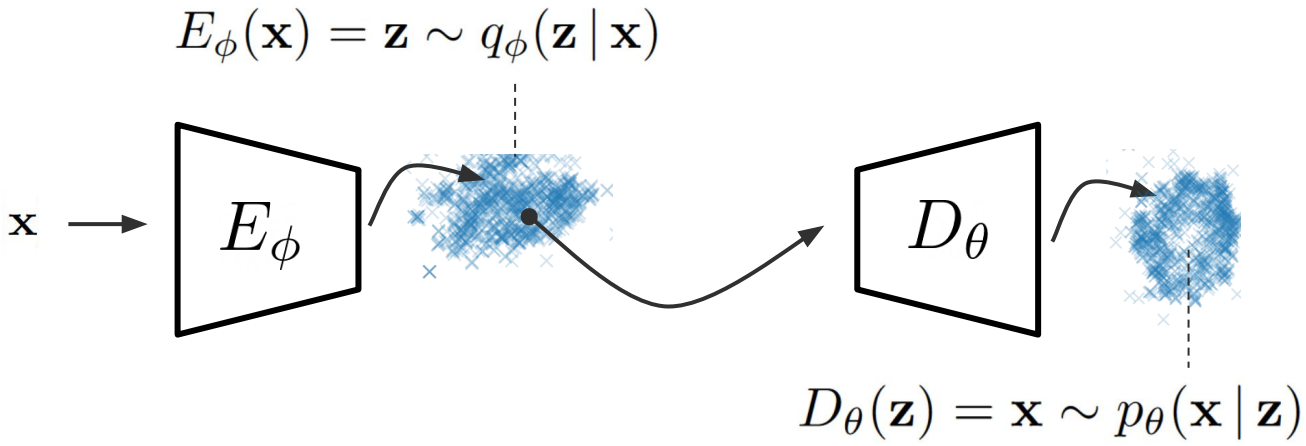
$$E_\phi(\mathbf{x}) = \mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$$



Variational Autoencoders (VAEs)

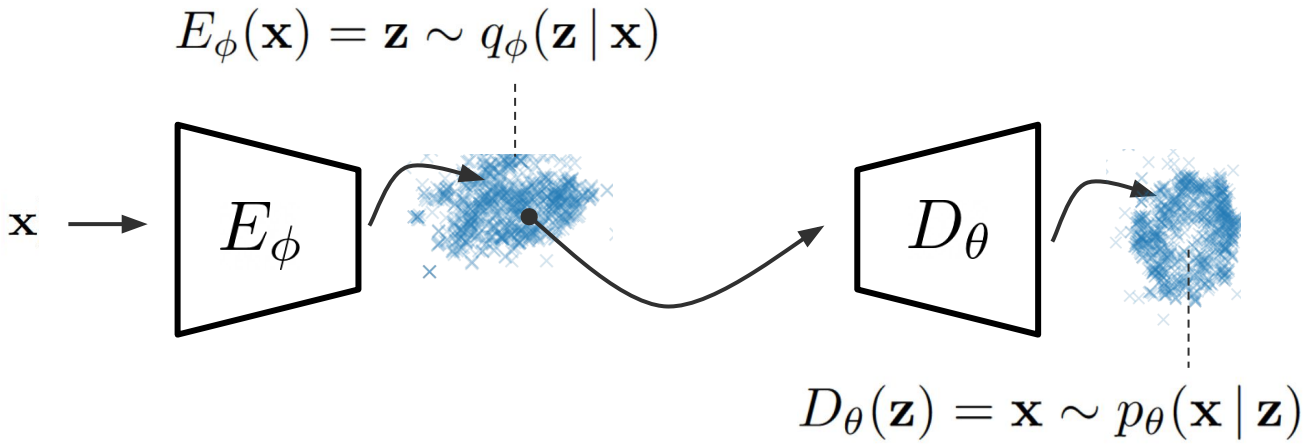


How to train VAEs?



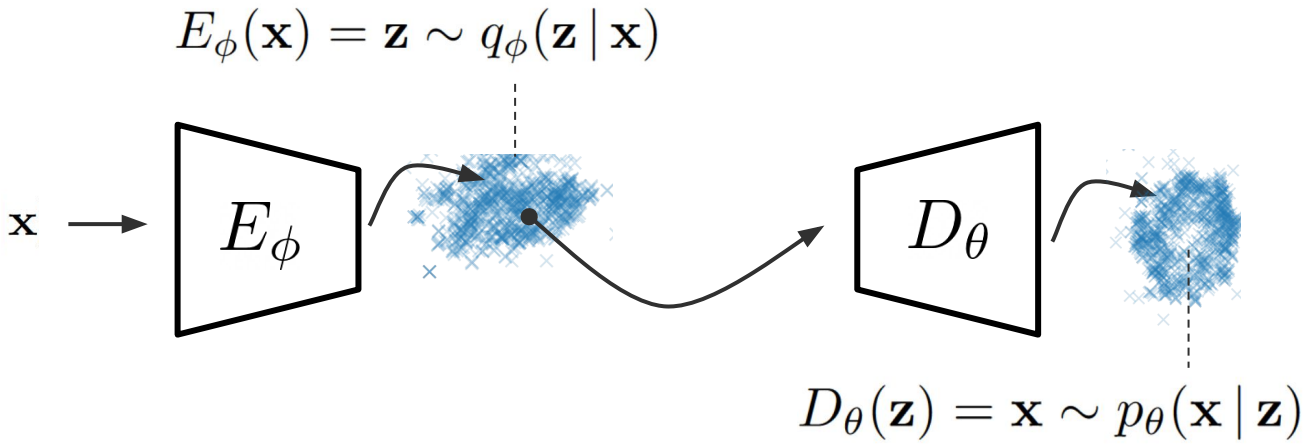
$$\log p_\theta(\mathbf{x}) \geq \text{ELBO}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) - \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))$$

How to train VAEs?



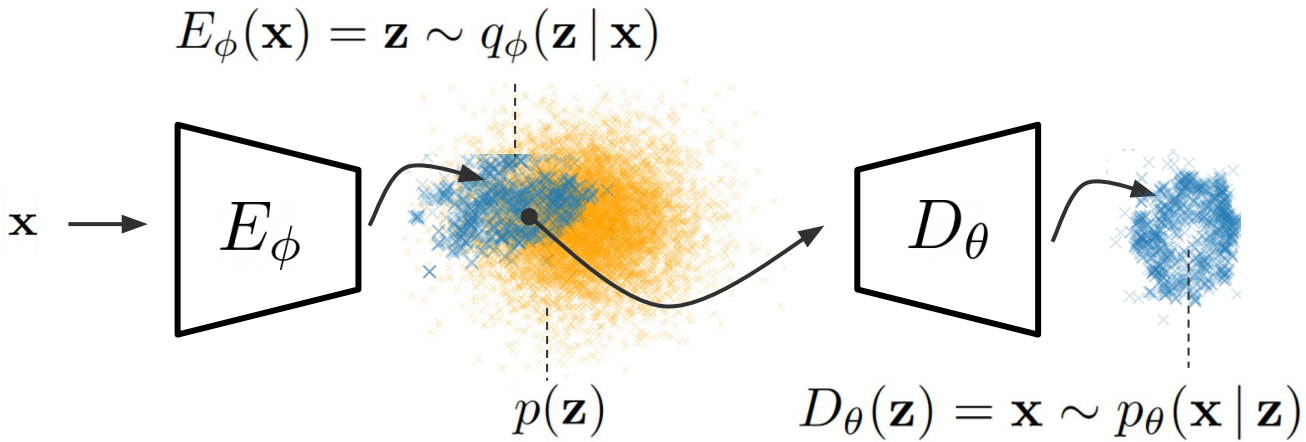
$$\log p_\theta(\mathbf{x}) \geq \mathsf{ELBO}(\phi, \theta, \mathbf{x}) = \underbrace{\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z})}_{\mathcal{L}_{\text{REC}}} - \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))$$

How to train VAEs?



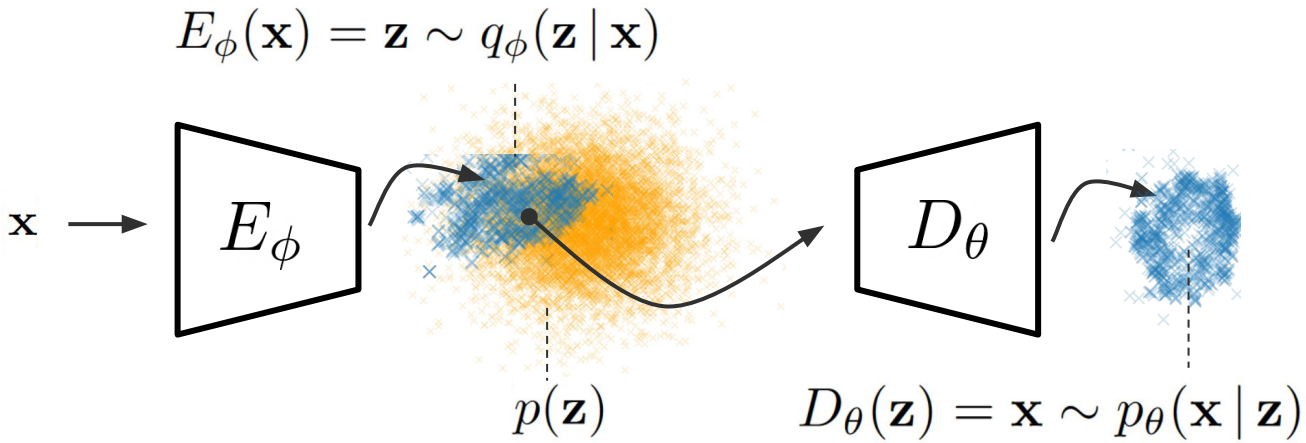
$$\log p_\theta(\mathbf{x}) \geq \mathsf{ELBO}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) - \underbrace{\mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))}_{\mathcal{L}_{\mathsf{KL}}}$$

How to train VAEs?



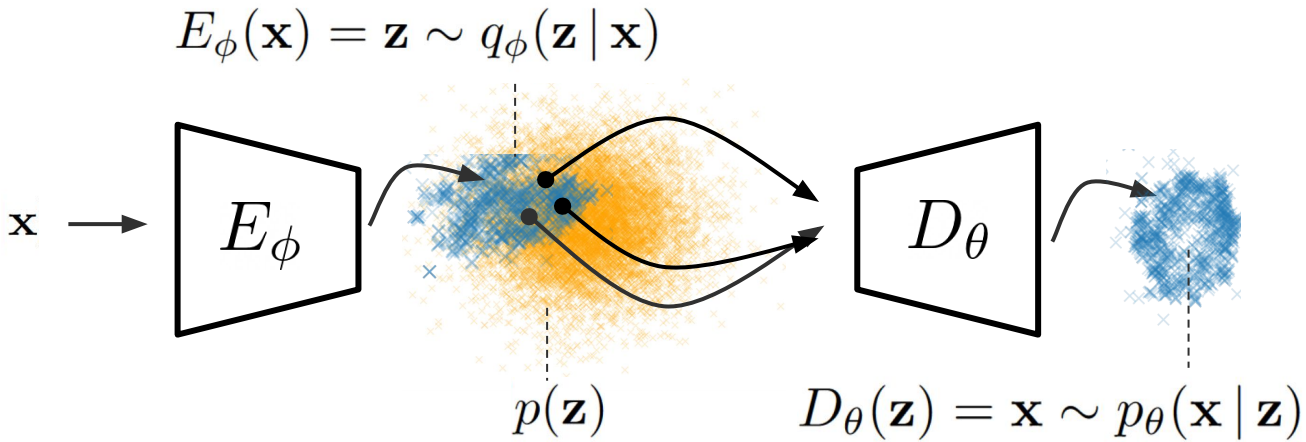
$$\log p_\theta(\mathbf{x}) \geq \text{ELBO}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) - \underbrace{\mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))}_{\mathcal{L}_{\text{KL}}}$$

Training VAEs: issues



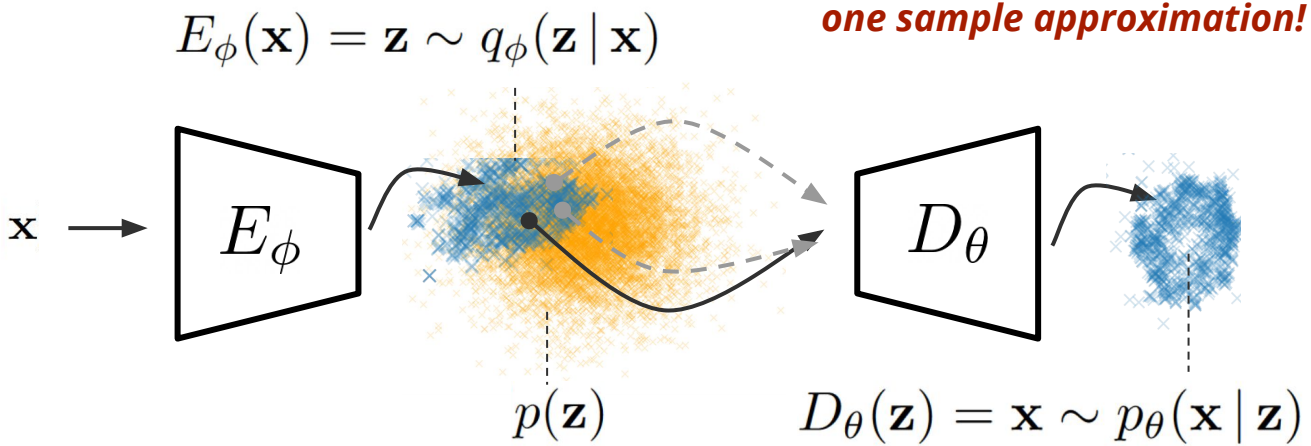
Balancing reconstruction quality and compression *[Burda et al. 2015, Tolshkin et al. 2018, ...]*
Spurious global optima *[Dai et al. 2019]*
Posterior collapse *[van den Oord et al. 2018, ...]*
Prior/aggregate posterior mismatch *[Tolshkin et al. 2018, Dai et al. 2019, ...]*

Issue #1: balancing training



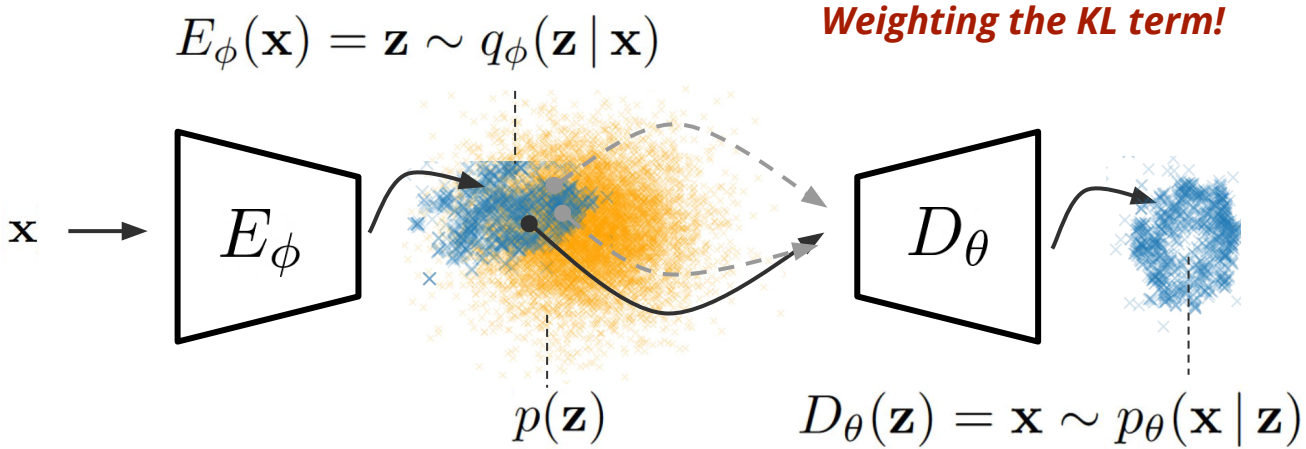
$$\log p_\theta(\mathbf{x}) \geq \text{ELBO}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) - \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))$$

Issue #1: balancing training



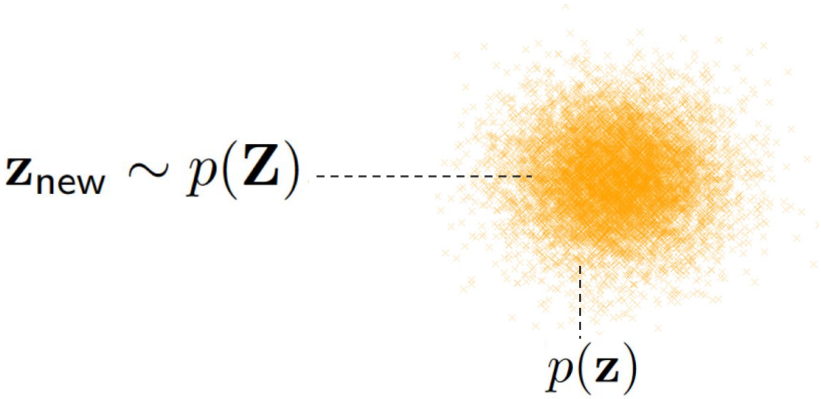
$$\log p_\theta(\mathbf{x}) \geq \text{ELBO}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) - \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))$$

Issue #1: balancing training

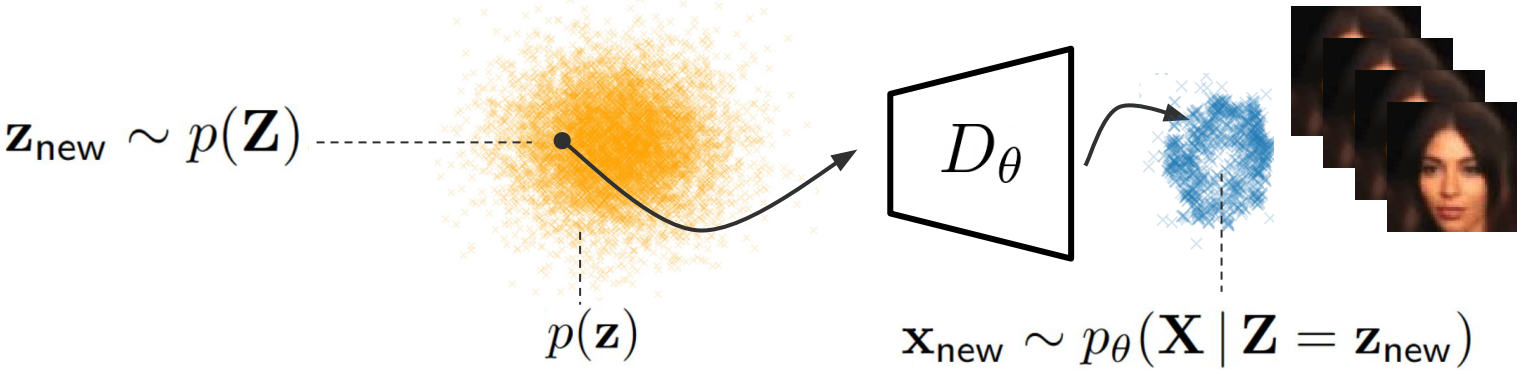


$$\log p_\theta(\mathbf{x}) \geq \text{ELBO}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) - \boxed{\mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))}$$

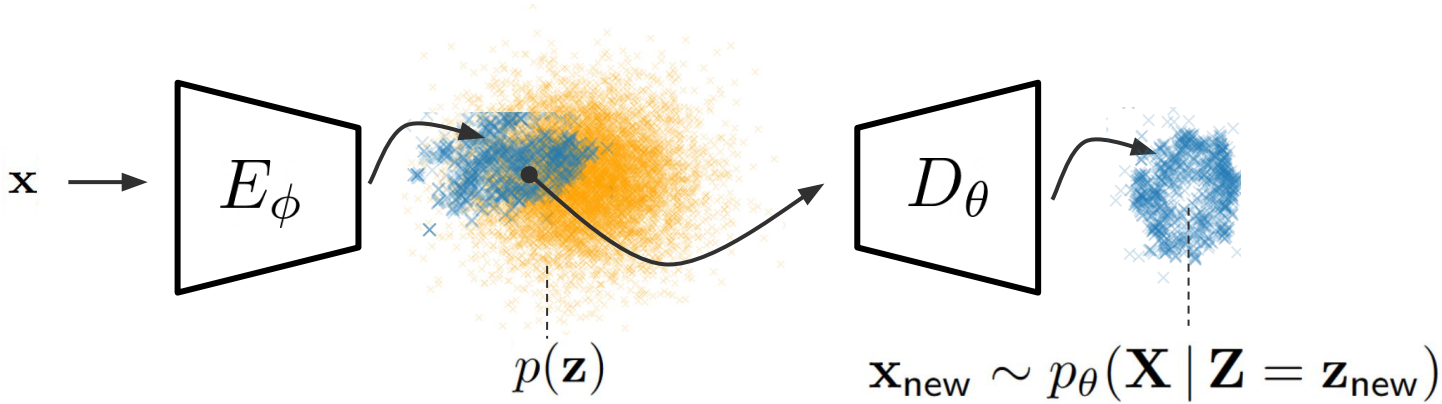
Sampling VAEs



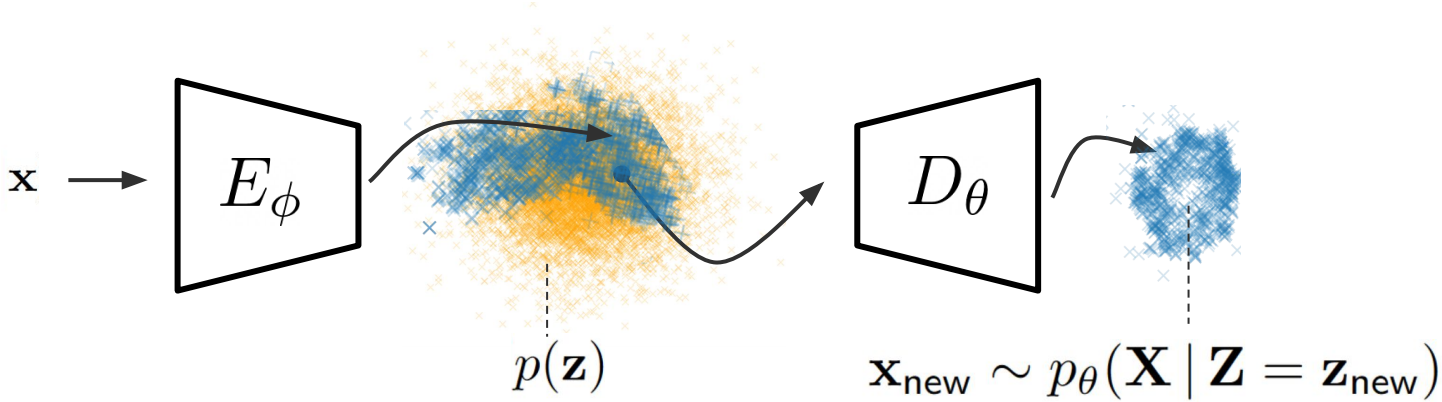
Sampling VAEs



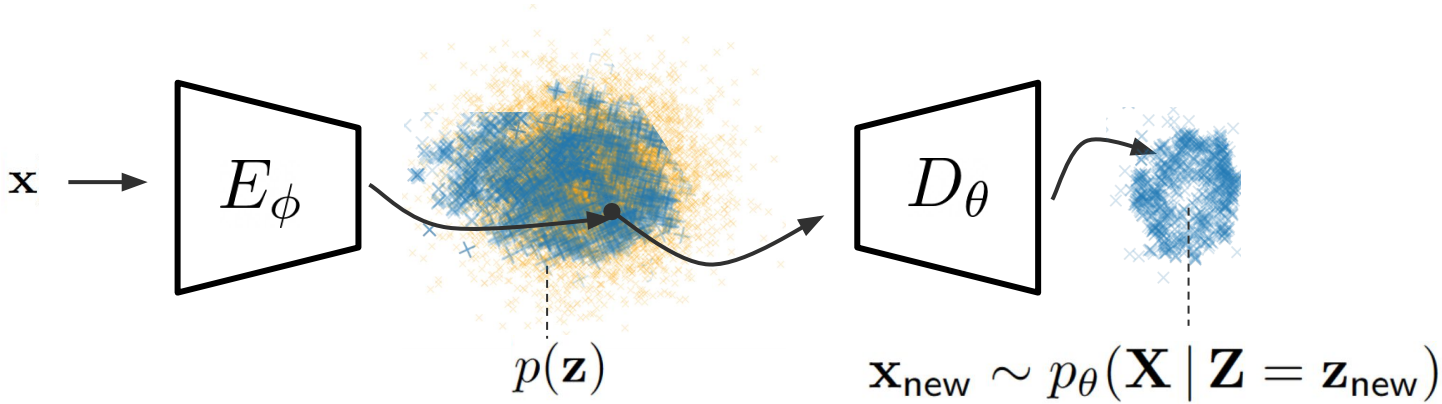
Sampling VAEs



Sampling VAEs

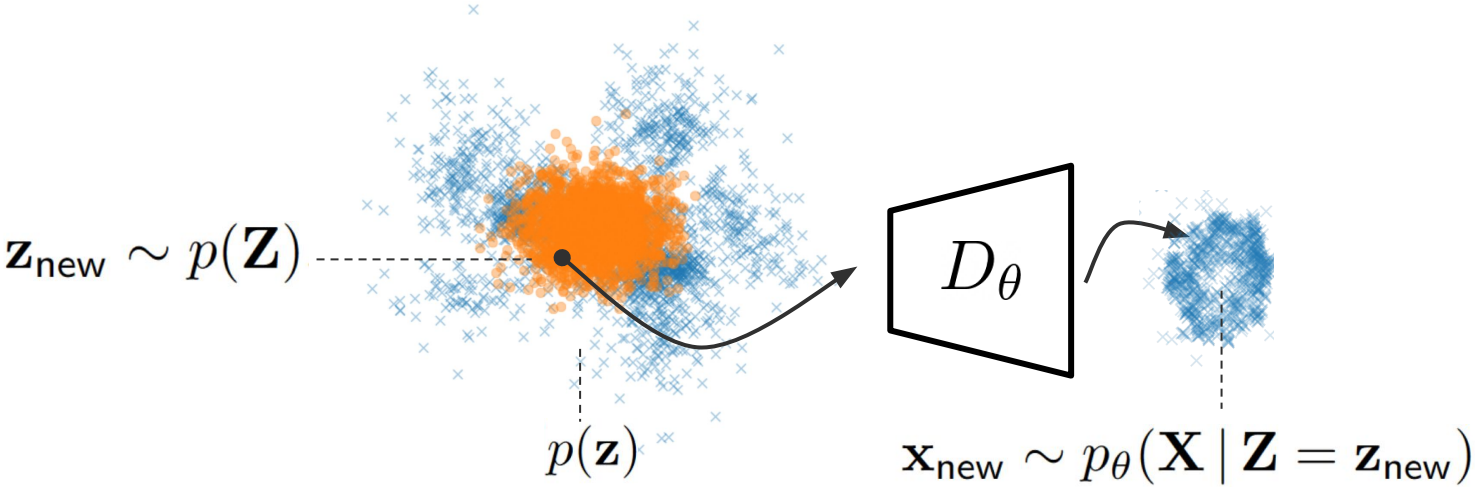


Sampling VAEs



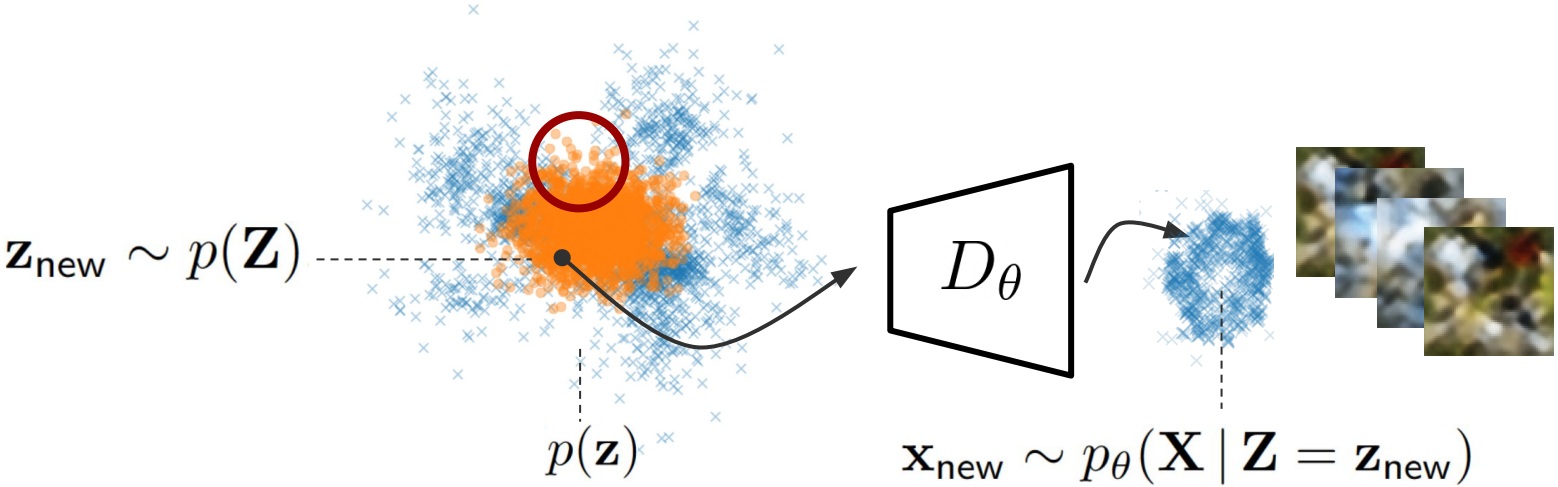
the aggregate posterior should ***ideally*** match the prior!

Issue #2: sampling spurious codes



the prior/aggregate posterior mismatch

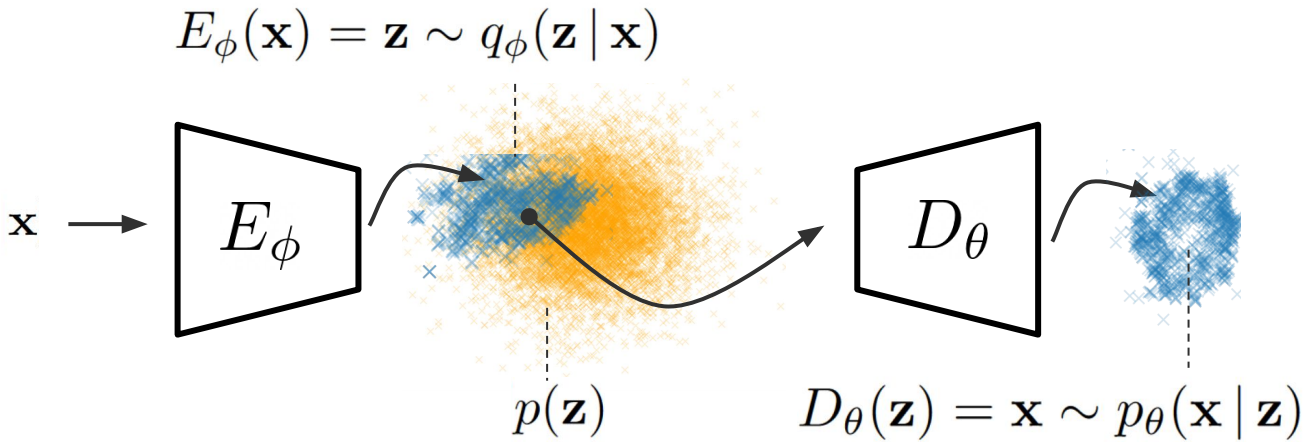
Issue #2: sampling spurious codes



the decoder has a hard time "imagining"

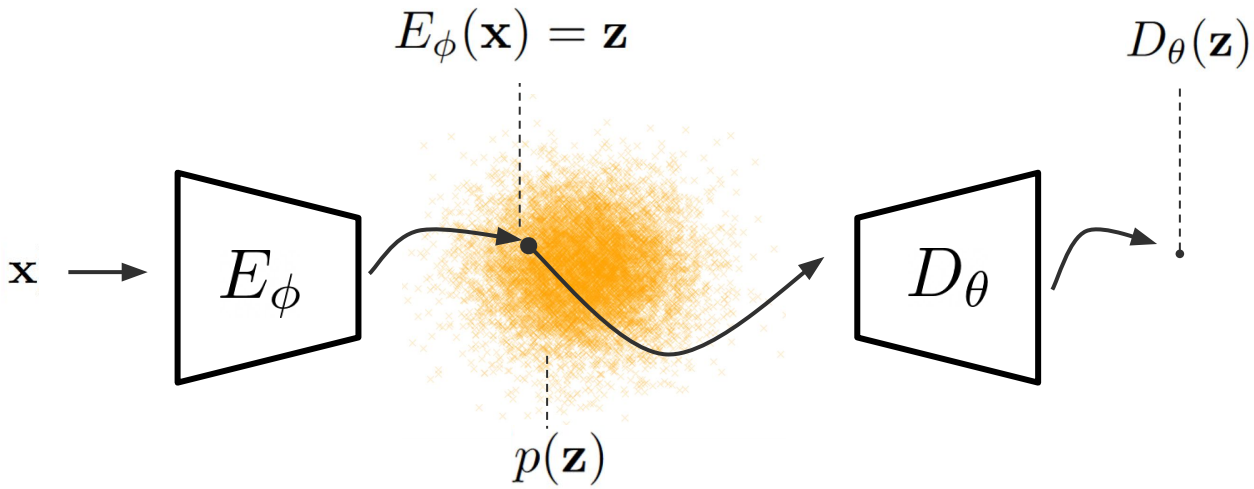
Can we do better?

Simpler VAEs?



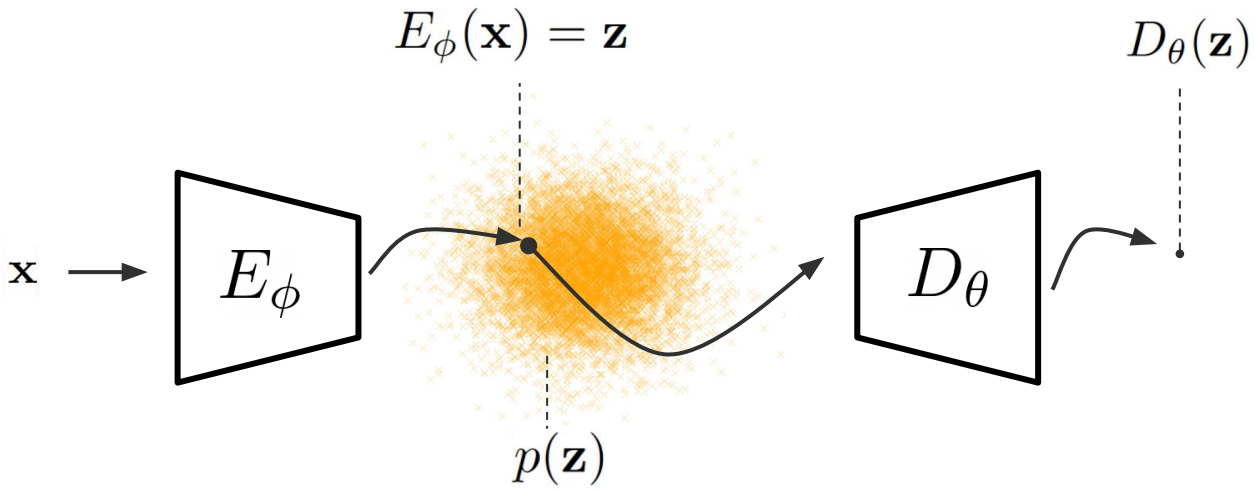
$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))$$

Simpler VAEs?



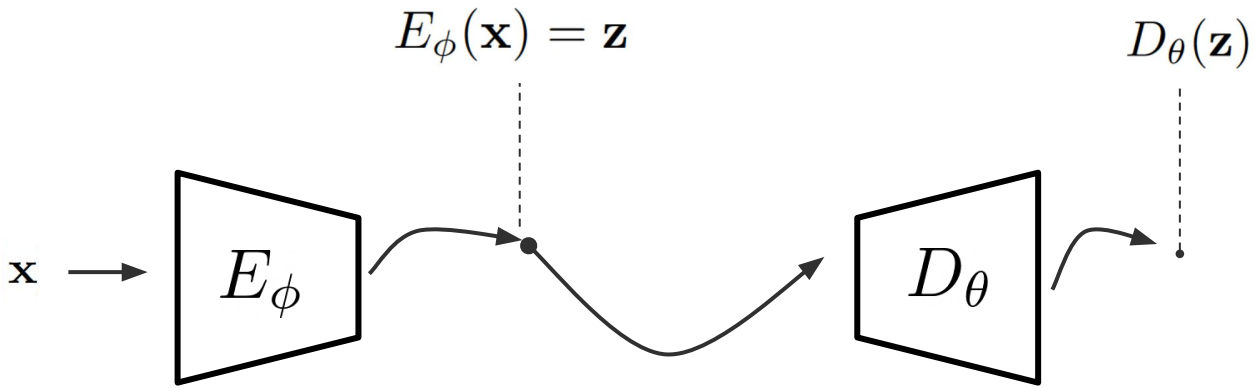
$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

Simpler VAEs?



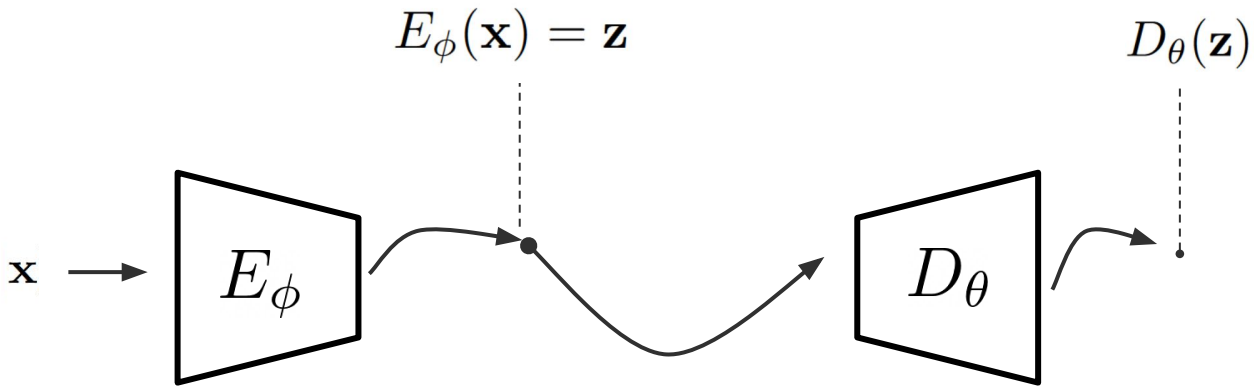
$$\mathcal{L}_{\text{VAE}} = \overbrace{||\mathbf{x} - D_\theta(\mathbf{z})||_2^2}^{\mathcal{L}_{\text{REC}}} + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

Simpler VAEs?



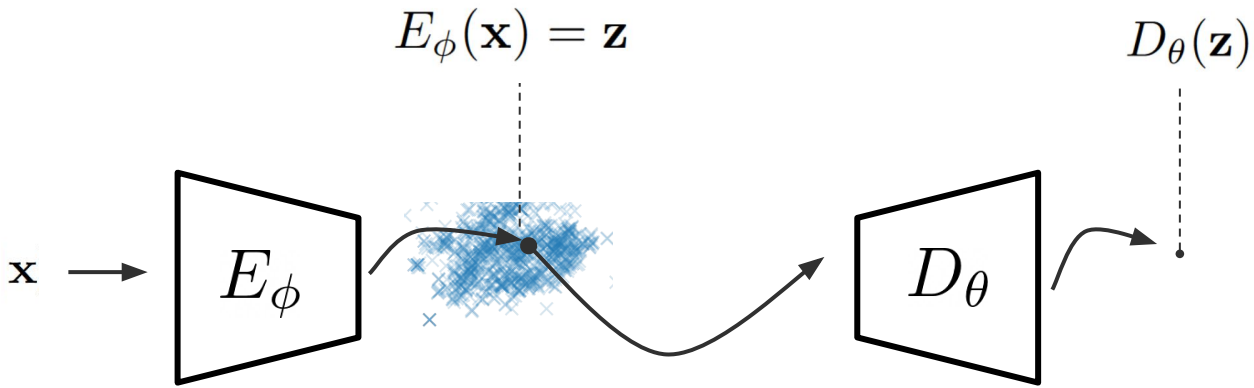
$$\mathcal{L}_{\text{VAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2$$

Simpler VAEs?



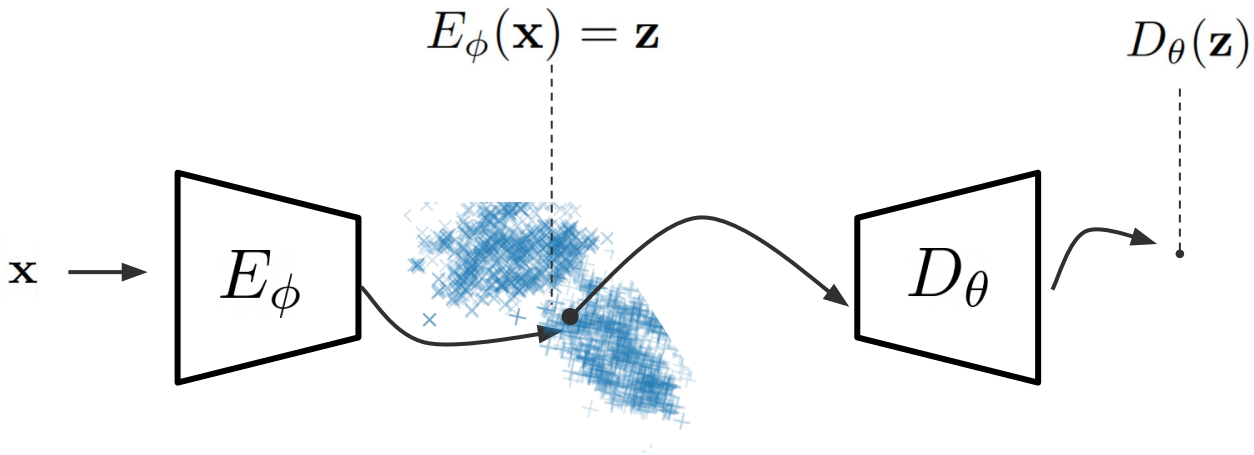
$$\mathcal{L}_{\text{VAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \beta ||\mathbf{z}||_2^2$$

Simpler VAEs?



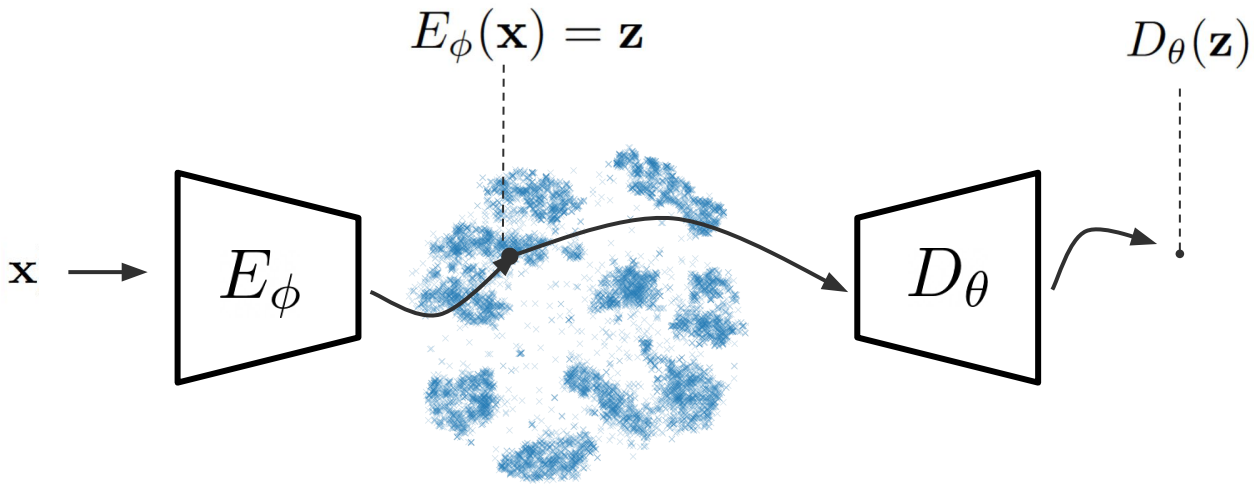
$$\mathcal{L}_{\text{VAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \beta ||\mathbf{z}||_2^2$$

Simpler VAEs?



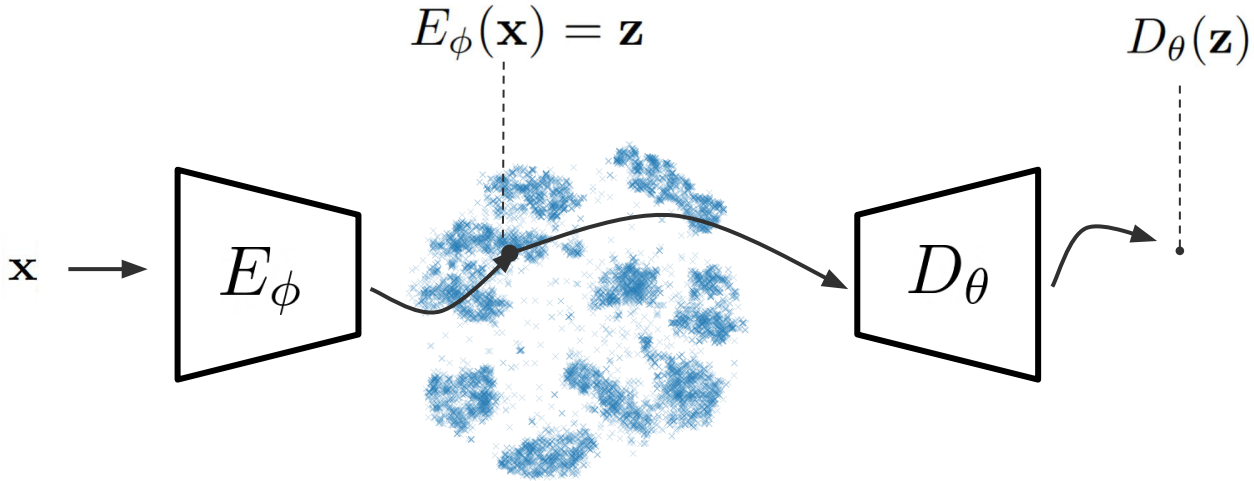
$$\mathcal{L}_{\text{VAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \beta ||\mathbf{z}||_2^2$$

Simpler VAEs?



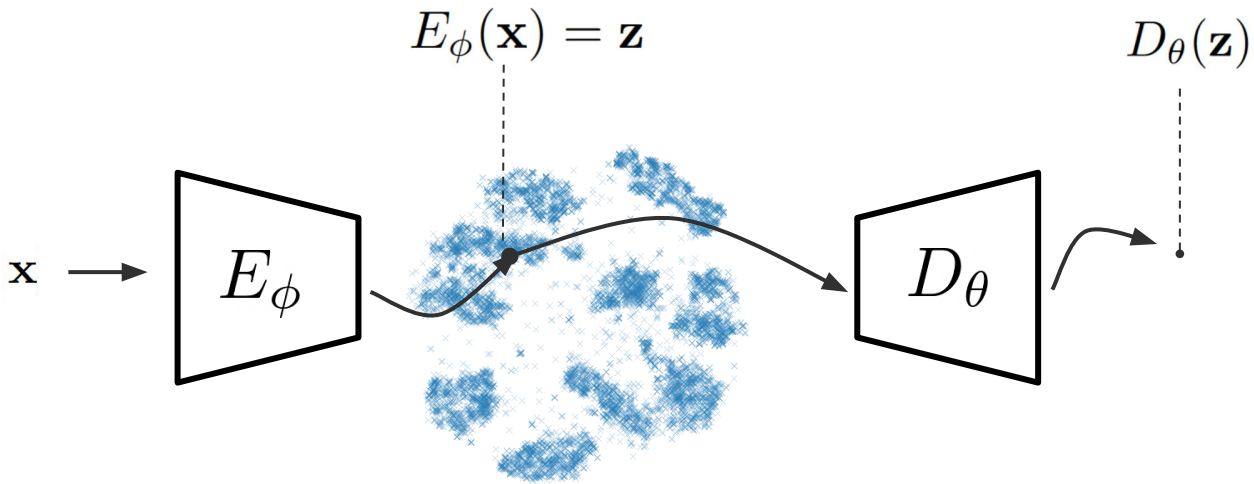
$$\mathcal{L}_{\text{VAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \beta ||\mathbf{z}||_2^2$$

How to have a smooth latent space?



ideally, $||\mathbf{z}_1 - \mathbf{z}_2||_p < \epsilon_1 \implies ||D_\theta(\mathbf{z}_1) - D_\theta(\mathbf{z}_2)||_p < \epsilon_2$

Regularized Autoencoders (RAEs)!



$$\mathcal{L}_{\text{RAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \beta ||\mathbf{z}||_2^2 + \lambda \mathcal{L}_{\text{REG}}$$

Which regularization for RAEs?

Which regularization for RAEs?

Gradient penalization *[Gulrajani et al. 2017; Mescheder et al. 2018]*

$$\mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{GP}} = ||\nabla D_\theta(E_\phi(\mathbf{x}))||_2^2$$

Which regularization for RAEs?

Gradient penalization *[Gulrajani et al. 2017; Mescheder et al. 2018]*

$$\mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{GP}} = ||\nabla D_\theta(E_\phi(\mathbf{x}))||_2^2$$

Spectral normalization *[Miyato et al. 2018]*

$$\theta_\ell^{\text{SN}} = \theta_\ell / \mathsf{s}(\theta_\ell)$$

Which regularization for RAEs?

Gradient penalization *[Gulrajani et al. 2017; Mescheder et al. 2018]*

$$\mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{GP}} = ||\nabla D_\theta(E_\phi(\mathbf{x}))||_2^2$$

Spectral normalization *[Miyato et al. 2018]*

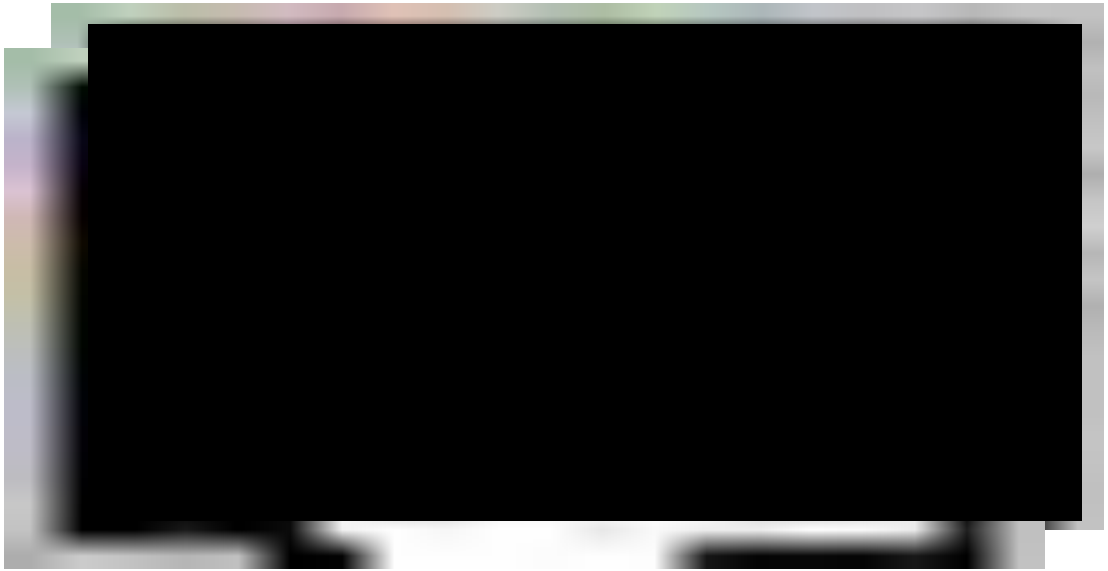
$$\theta_\ell^{\text{SN}} = \theta_\ell / \text{s}(\theta_\ell)$$

Weight decay *[Bishop et al. 1996]*

$$\mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{L}_2} = ||\theta||_2^2$$

RAE for image generation

VAE RAE+*L2*

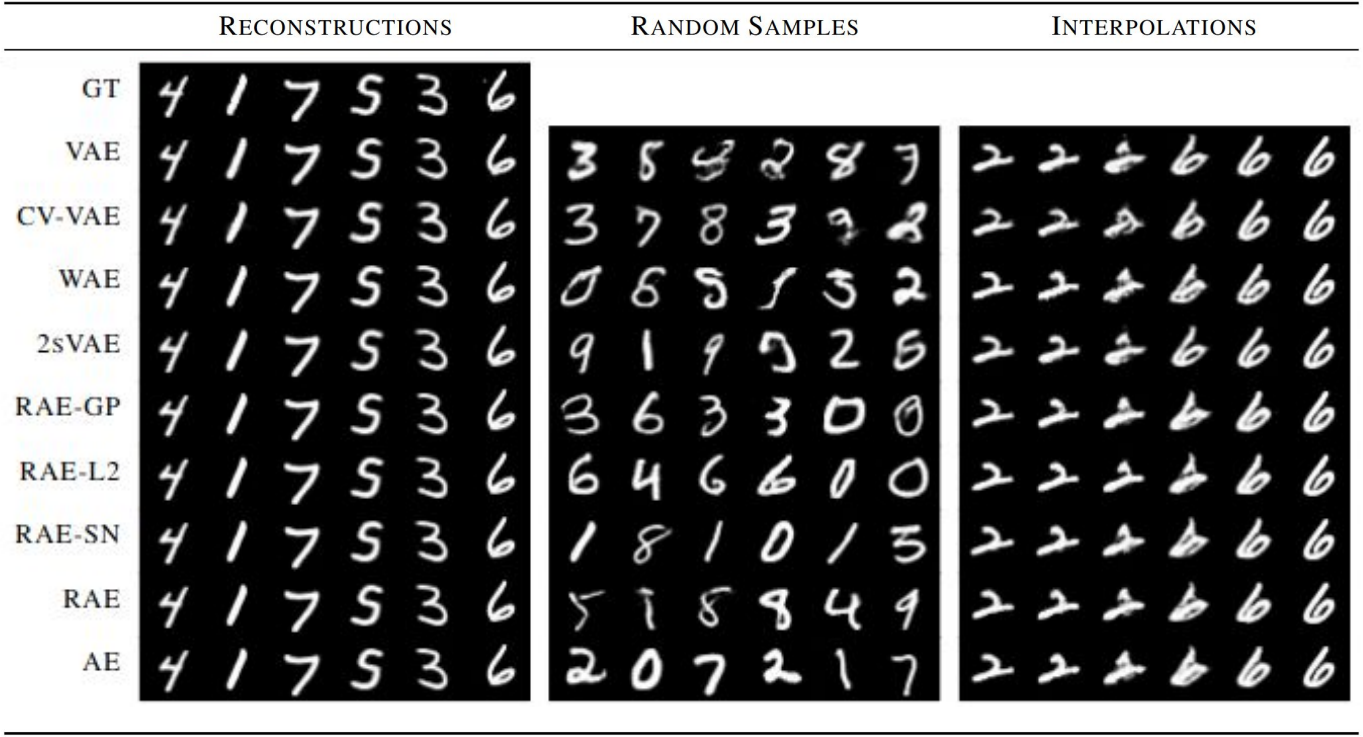


RAEs generate ***equally good or better samples and interpolations***

RAE for image generation

*even when regularization is **implicit!***

Common image benchmarks: MNIST

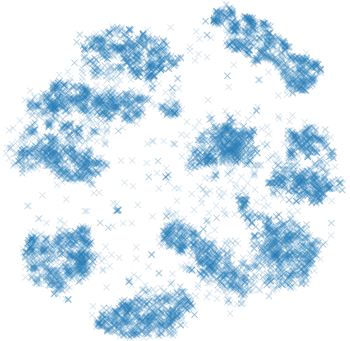


Common image benchmarks: CIFAR10

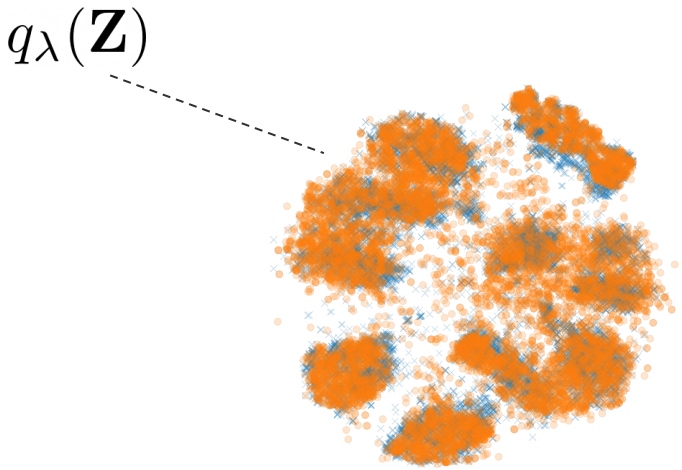
Common image benchmarks: CelebA

How do we sample from RAEs...?

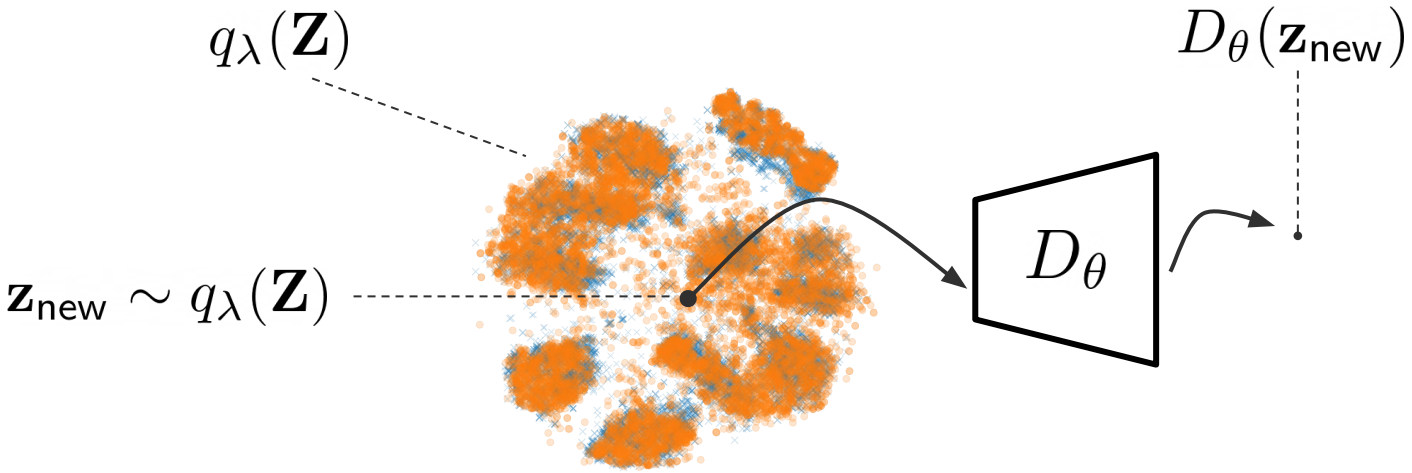
Sampling RAEs…?



Ex-Post Density Estimation (XPDE)



Ex-Post Density Estimation (XPDE)



Which density estimator for XPDE?

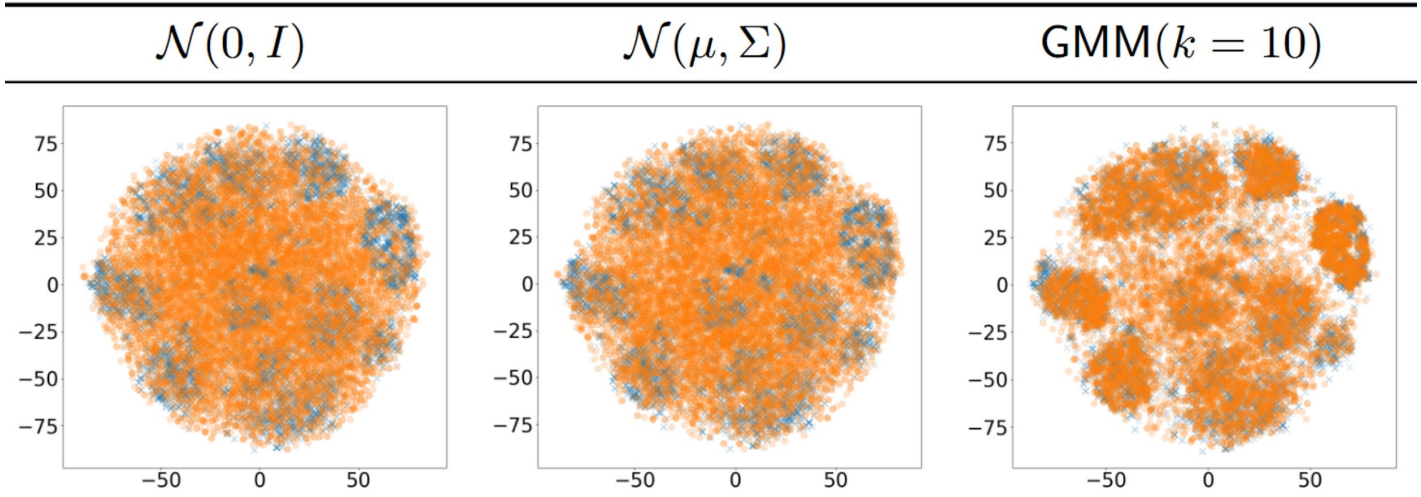
Which density estimator for XPDE?

a SOTA deep generative model *e.g. autoregressive model or Flow*
[van den Oord et al. 2019, Razavi et al. 2020]

...or another VAE!

- ⇒ ***VAE training and sampling issues ...are still there!***

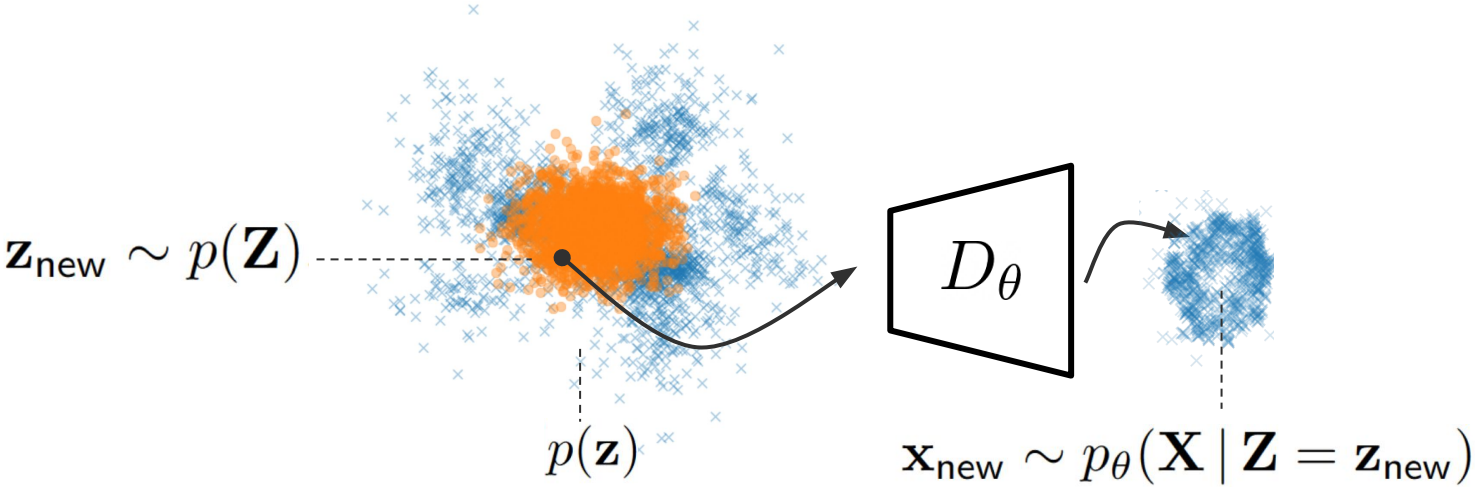
Which density estimator for XPDE?



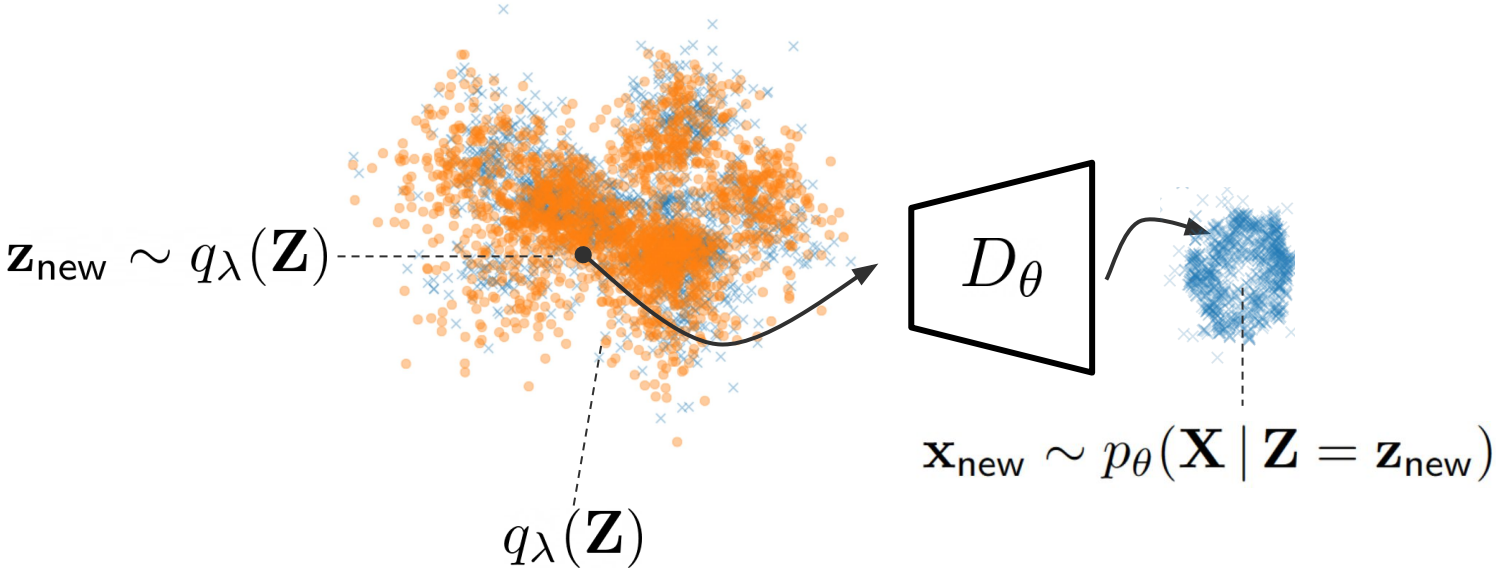
striving for simplicity: just ***Gaussian Mixture Models***

Can’t we just do XPDE for VAEs?

Can’t we just do XPDE for VAEs?



Can’t we just do XPDE for VAEs?



Ex-Post Density Estimation (XPDE)

	MNIST		CIFAR		CelebA	
	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM
VAE	19.21	**17.66**	106.37	**103.78**	48.12	**45.52**
CV-VAE	33.79	**17.87**	94.75	**86.64**	48.87	**49.30**
WAE	20.42	**9.39**	117.44	**93.53**	53.67	**42.73**
RAE-GP	22.21	**11.54**	83.05	**76.33**	116.30	**45.63**
RAE-L2	22.22	**8.69**	80.80	**74.16**	51.13	**47.97**
RAE-SN	19.67	**11.74**	84.25	**75.30**	44.74	**40.95**
RAE	23.92	**9.81**	83.87	**76.28**	48.20	**44.68**
AE	58.73	**10.66**	84.74	**76.47**	127.85	**45.10**
AE-L2	315.15	**9.36**	247.48	**75.40**	346.29	**48.42**

*XPDE consistently improves sample quality **for all VAE variants***

Why…does it work?

Why…does it work?

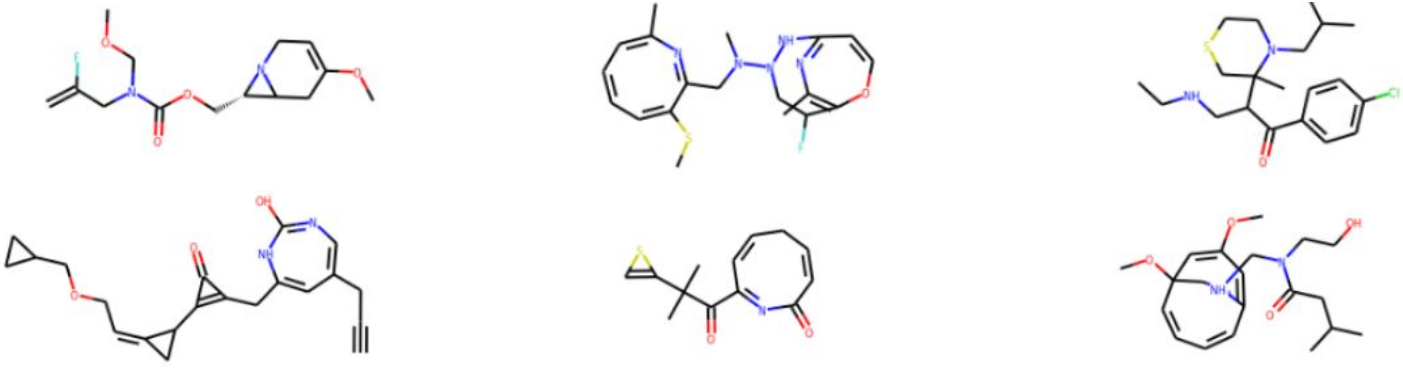
ConvNets are very, very, very ***smooth***! *[LeCun et al. 1994]*

Why…does it work?

ConvNets are very, very, very ***smooth***! *[LeCun et al. 1994]*

…and these datasets are full, full, ***full of regularities***!

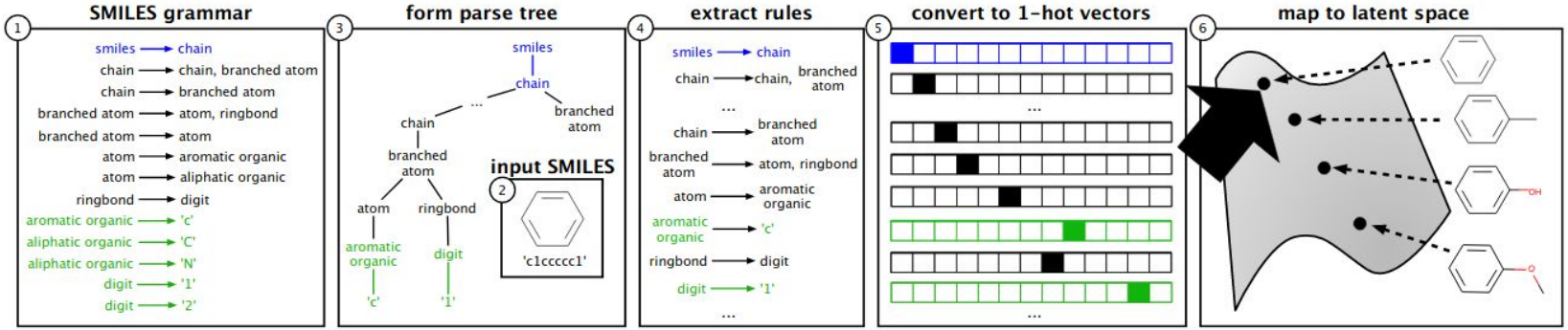
What about more challenging data?



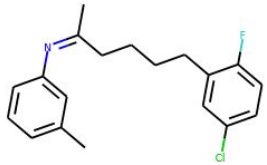
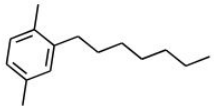
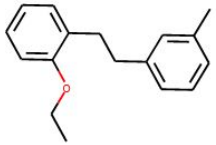
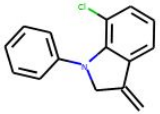
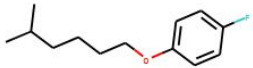
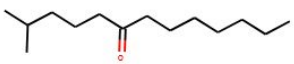
E.g., generating structured objects like molecules

VAEs for molecules?

- ⇒ Molecule VAE *[Bombardelli et al. 2017]*
- ⇒ ***GrammarVAE (GVAE)*** *[Kusner et al. 2019]*
- ⇒ Constrained Graph VAE (CGVAE) *[Liu et al. 2018, ...]*
- ⇒ **...**



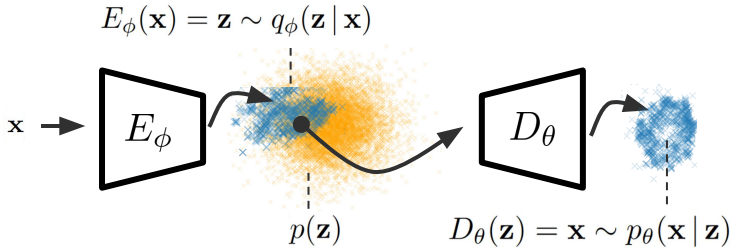
GRAE: RAEifying the Grammar VAE

GRAE			
SCORE	**3.74**	**3.52**	**3.14**
GVAE			
SCORE	3.13	3.10	2.37

More accurate generation than *Kusner et al. 2017*

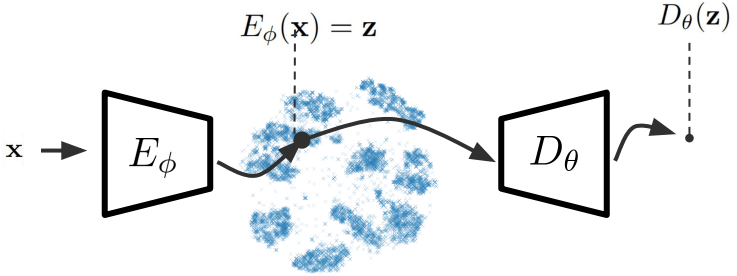
RAEify your VAEs!

VAE



$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

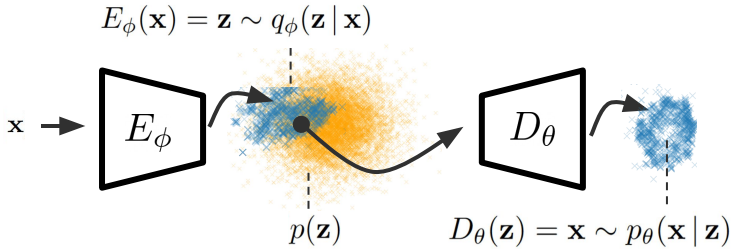
RAE



$$\mathcal{L}_{\text{RAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

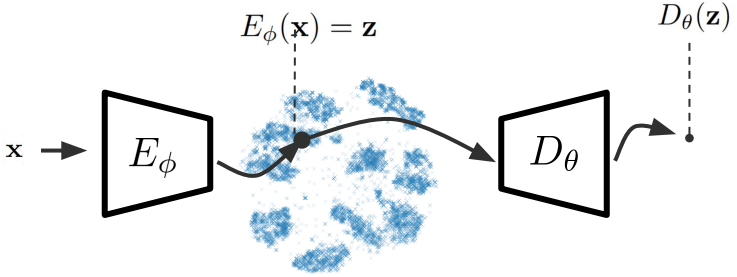
RAEify your VAEs!

VAE



$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

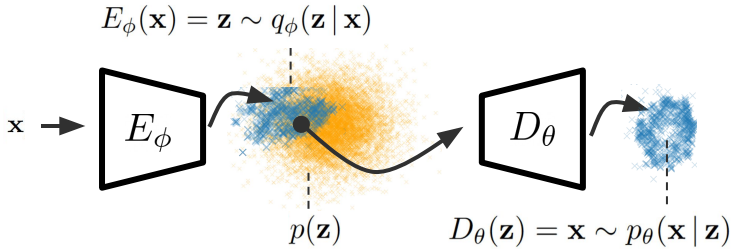
RAE



$$\mathcal{L}_{\text{RAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

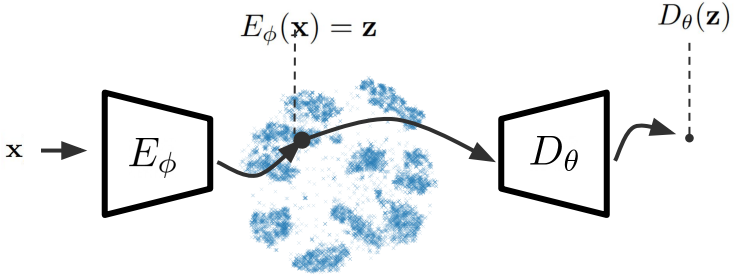
RAEify your VAEs!

VAE



$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

RAE



$$\mathcal{L}_{\text{RAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \beta\,||\mathbf{z}||_2^2 + \lambda\mathcal{L}_{\text{REG}}$$

Is this really simple... and new?

AEs for generative modeling

A Generative Process for Sampling Contractive Auto-Encoders

Salah Rifai[1]
Yoshua Bengio[1]
Yann N. Dauphin[1]
Pascal Vincent[1]
[1] Dept. IRO, Université de Montréal. Montréal (QC), H3C 3J7

Generalized Denoising Auto-Encoders as Generative Models

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent
Département d'informatique et recherche opérationnelle, Université de Montréal

MCMC schemes to sample from Contractive *[Rifai et al. 2011]*
and Denoising Autoencoders *[Bengio et al. 2009]*

Other flavours of XPDE

Two-Stage VAEs *[Dai et al. 2019]*

use another VAE for XPDE

⇒ *VAE training and sampling issues …are still there!*

Diagnosing and Enhancing VAE Models

Bin Dai DAIB13@MAILS.TSINGHUA.EDU.CN
Institute for Advanced Study
Tsinghua University
Beijing, China

David Wipf DAVIDWIPF@GMAIL.COM
Microsoft Research
Beijing, China

VQ-VAEs *[van den Oord et al. 2019, Razavi et al. 2020]*

use PixelCNN over discrete latents

⇒ *VQ-VAEs are RAEs not VAEs!*

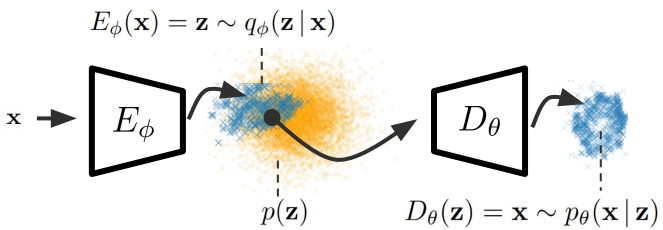
Neural Discrete Representation Learning

Aaron van den Oord DeepMind avdnoord@google.com
Oriol Vinyals DeepMind vinyals@google.com
Koray Kavukcuoglu DeepMind korayk@google.com

What did we lose?

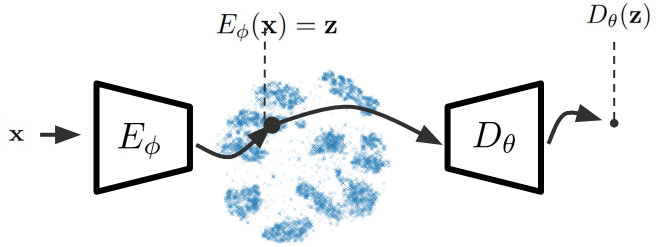
What did we lose?

Variational Autoencoders
(VAEs)



⇒ Generative modeling ✓
⇒ Density Estimation ✓
⇒ Disentanglement ✓

Regularized Autoencoders
(RAEs)



⇒ Generative modeling ✓
⇒ Density Estimation ***?***
⇒ Disentanglement ***?***

RAEs for density estimation ?

RAEs (and VQ-VAEs) are like GANs, they are ***implicit likelihood models!***

~~$\log p_\theta(\mathbf{x}) \geq \text{ELBO}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) \quad \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z}))$~~

RAEs for density estimation (?)

RAEs (and VQ-VAEs) are like GANs, they are ***implicit likelihood models!***

An approximate ***ELBO can be recovered*** under some geometric assumptions

---

Regularized Autoencoders via Relaxed Injective Probability Flow

---

Abhishek Kumar
Google Research

Ben Poole
Google Research

Kevin Murphy
Google Research

Flows for simultaneous manifold learning and density estimation

Johann Brehmer[a,b,1] **and Kyle Cranmer**[a,b]

[a]Center for Data Science, New York University, USA; [b]Center for Cosmology and Particle Physics, New York University, USA

RAEs for disentanglement (?)

DISENTANGLED REPRESENTATION LEARNING AND GENERATION WITH MANIFOLD OPTIMIZATION

Arun Pandey
Department of Electrical Engineering
ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
arun.pandey@esat.kuleuven.be

Michaël Fanuel
Department of Electrical Engineering
ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
michael.fanuel@esat.kuleuven.be

Variance Constrained Autoencoding

D. T. Braithwaite,* M. O'Connor, W. B. Kleijn
School of Engineering and Computer Science,
Victoria University of Wellington,
New Zealand

Conclusions

aiPhones

⇒ ***Phone capabilities***
⇒ aiCloud, aiWatch, aiTunes,...
⇒ 4k Video, ...

aiPhones

⇒ ***Phone capabilities***
⇒ aiCloud, aiWatch, aiTunes,...
⇒ 4k Video, ...

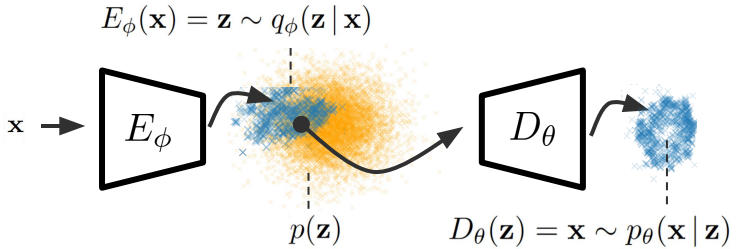
$\gtrsim$

RegularPhone

what is the simplest model that gets you further?

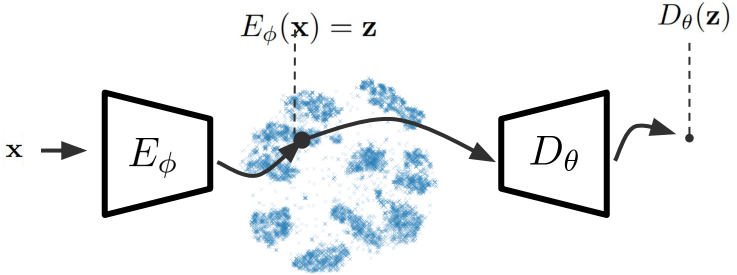
Takeaway #1: RAEify your VAEs!

VAE



$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + \mathbb{KL}(q_\phi(\mathbf{z} \mid \mathbf{x})||p(\mathbf{z}))$$

RAE



$$\mathcal{L}_{\text{RAE}} = ||\mathbf{x} - D_\theta(\mathbf{z})||_2^2 + \beta\,||\mathbf{z}||_2^2 + \lambda\mathcal{L}_{\text{REG}}$$

Takeaway #2: use XPDE!

	MNIST		CIFAR		CelebA	
	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM
VAE	19.21	**17.66**	106.37	**103.78**	48.12	**45.52**
CV-VAE	33.79	**17.87**	94.75	**86.64**	48.87	**49.30**
WAE	20.42	**9.39**	117.44	**93.53**	53.67	**42.73**
RAE-GP	22.21	**11.54**	83.05	**76.33**	116.30	**45.63**
RAE-L2	22.22	**8.69**	80.80	**74.16**	51.13	**47.97**
RAE-SN	19.67	**11.74**	84.25	**75.30**	44.74	**40.95**
RAE	23.92	**9.81**	83.87	**76.28**	48.20	**44.68**
AE	58.73	**10.66**	84.74	**76.47**	127.85	**45.10**
AE-L2	315.15	**9.36**	247.48	**75.40**	346.29	**48.42**

***Boost your VAEs** by training a density estimator on the latent codes!*

Paper

https://openreview.net/forum?id=S1g7tpEYDS

Code

https://github.com/ParthaEth/Regularized_autoencoders-RAE-