Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

09/10/2018

320 Soda Hall, University of California at Berkeley

CS294-150: Machine Learning and Statistics Meet Biology

Ryan Chung, Giulia Guidi, Weston Hughes, Hector Roux de Bézieux

Berkeley UNIVERSITY OF CALIFORNIA

Outline

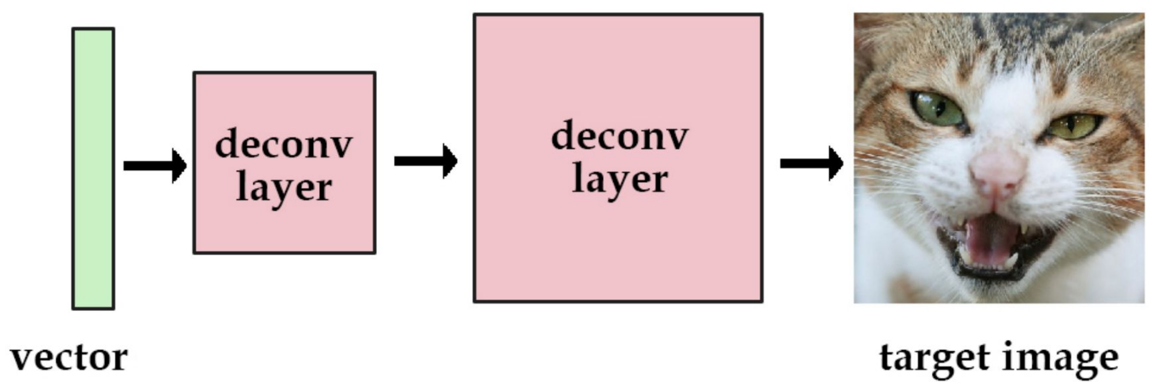
Introduction

Variational Autoencoders (VAEs):
The deep learning perspective

(Deep) generative models: We have some data, and we want to make more data following the same distribution

Also want to intelligently make new data by looking at where old data lies in a latent space

Two main methods: VAEs (2013) and GANs (2014)

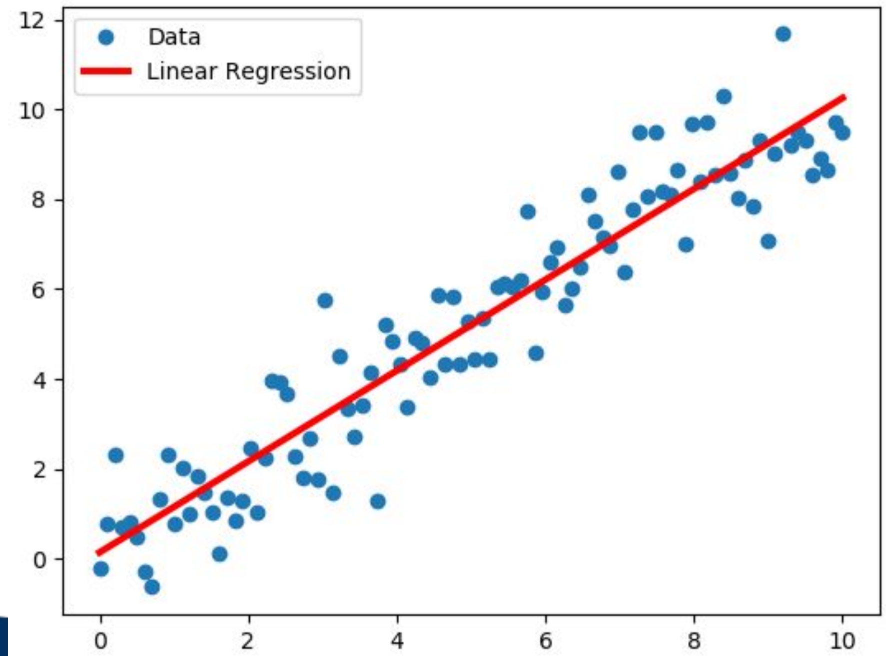


Berkeley
UNIVERSITY OF CALIFORNIA

Variational Autoencoders (VAEs):

The deep learning perspective

The Manifold Hypothesis: many data in high dimensional spaces lie in/near lower dimensional "manifolds"



Variational Autoencoders (VAEs):

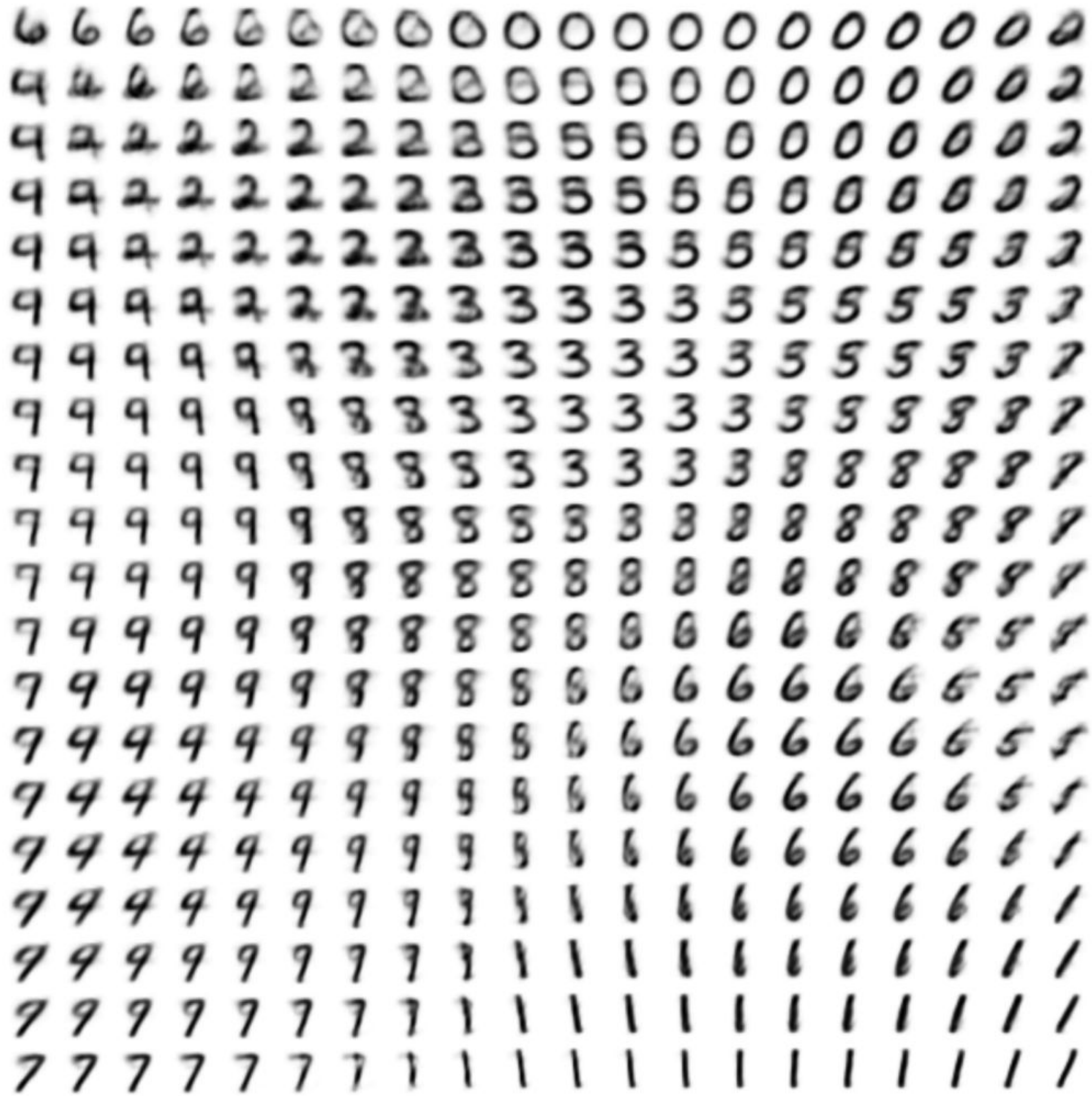
The deep learning perspective

MNIST: The space of numbers drawn in a 28x28 grid is a contiguous subset of the space of possible images

784 dimensional images, but data only exist in “small area” of space

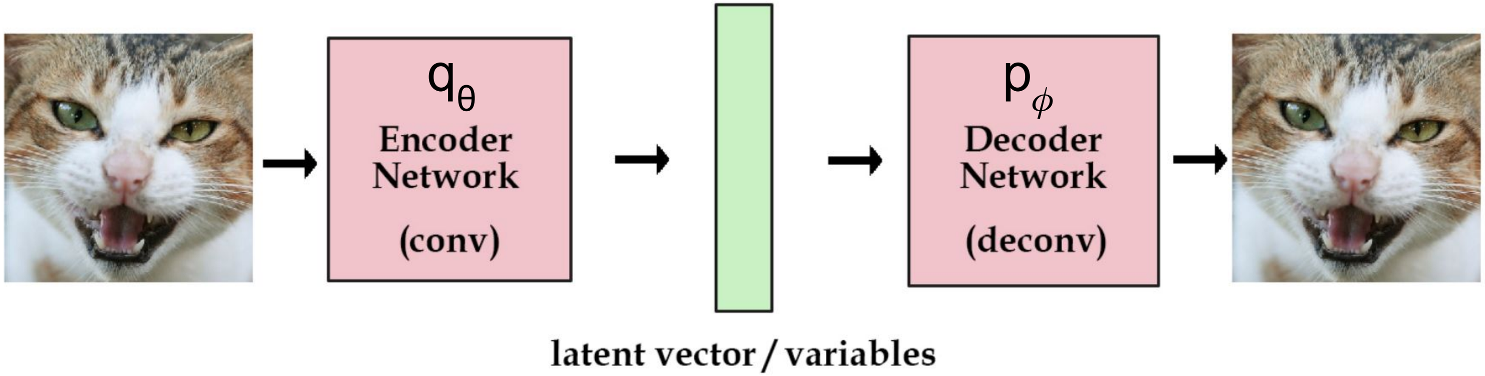
We assume this space is continuous

Can we use a neural network to find a low dimensional description of this “latent” space?



Variational Autoencoders (VAEs):
The deep learning perspective

First pass: information bottleneck through a lower dimensional latent space z

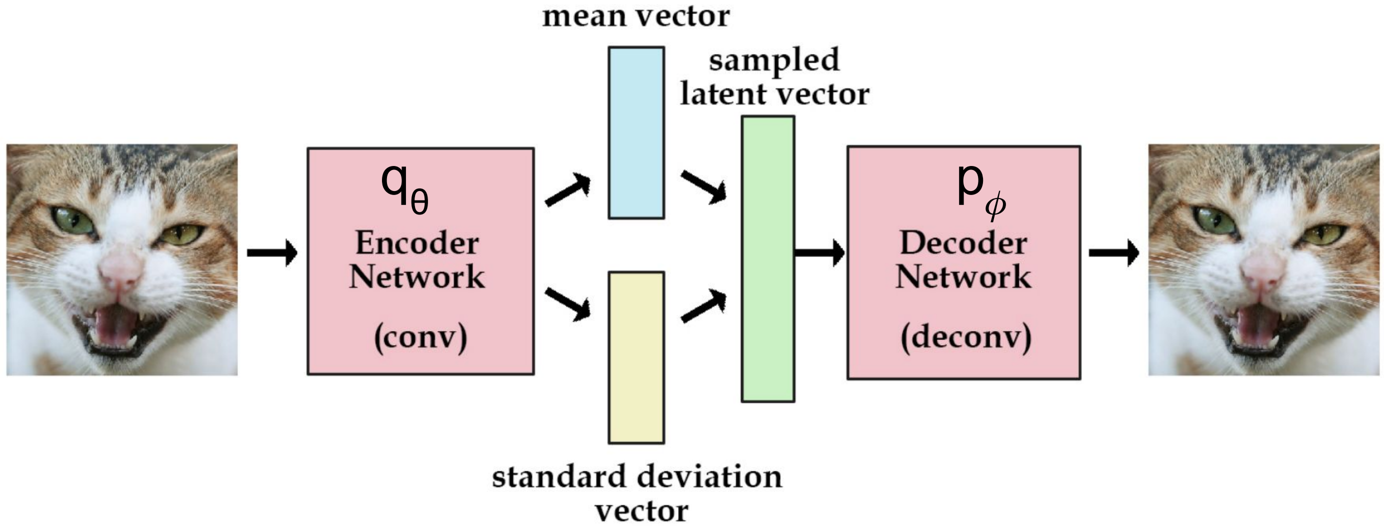


$$l_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)]$$

Variational Autoencoders (VAEs):

The deep learning perspective

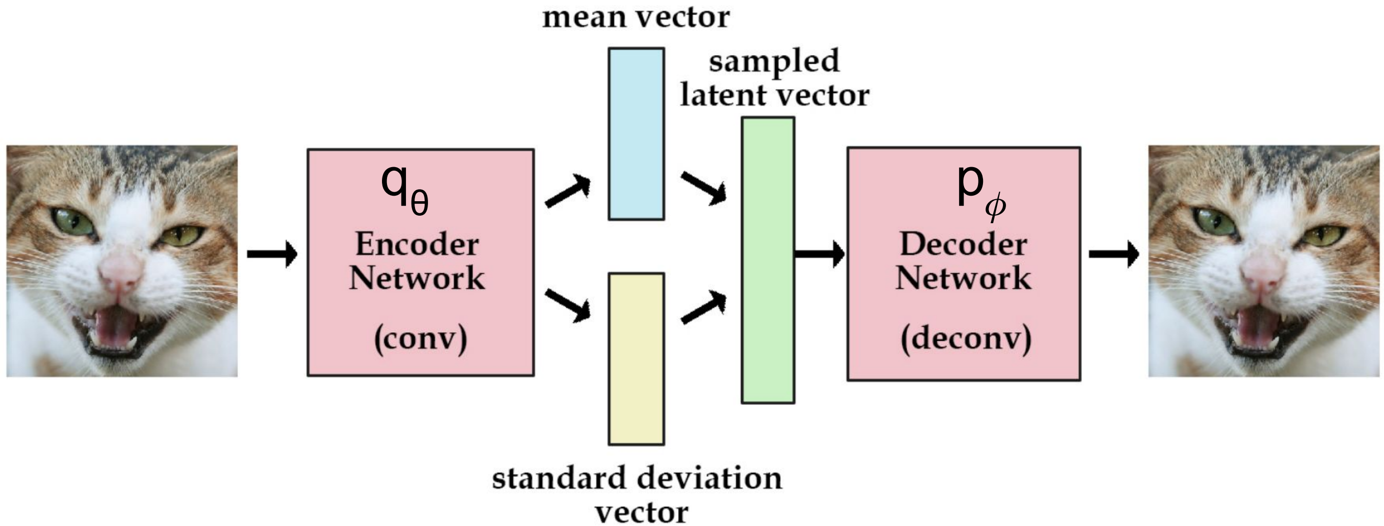
Second pass: add Gaussian stochasticity



Variational Autoencoders (VAEs):

The deep learning perspective

Third pass: regularize q to output distributions similar to standard normals

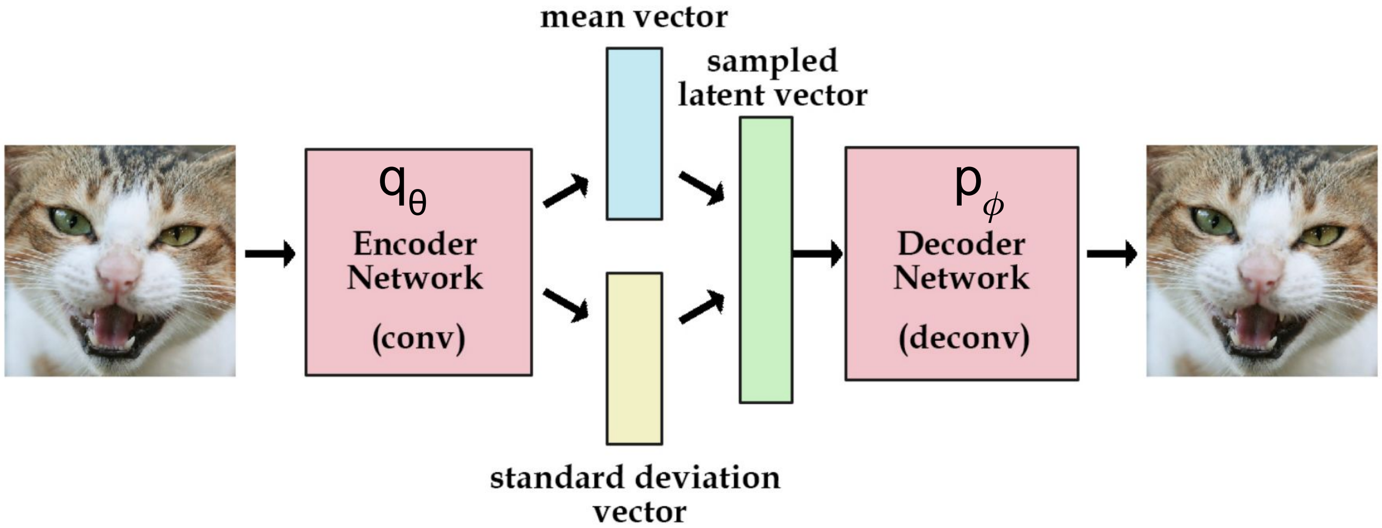


$$KL(q_\theta(z|x_i)||p(z))$$

Berkeley
UNIVERSITY OF CALIFORNIA

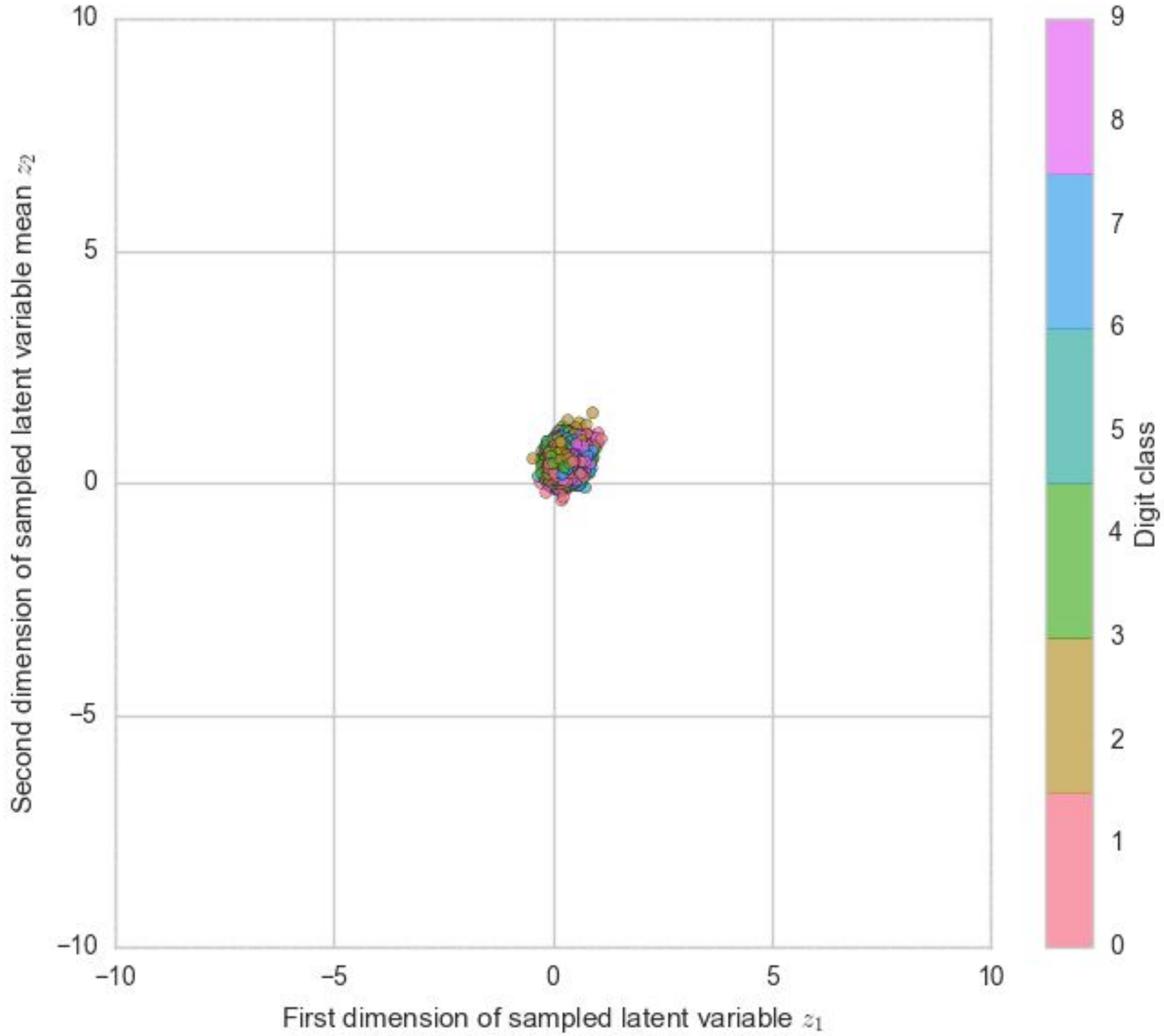
Variational Autoencoders (VAEs):

The deep learning perspective



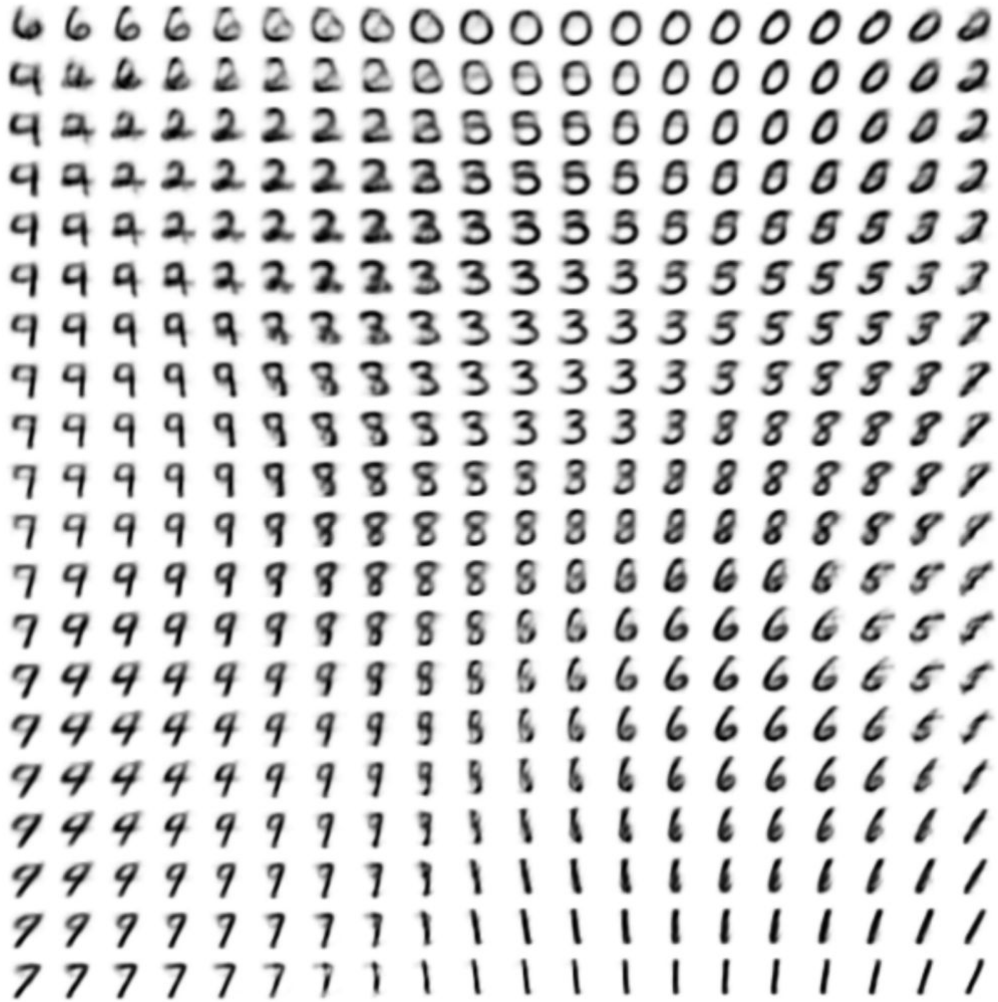
$$l_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + KL(q_\theta(z|x_i)||p(z))$$

Berkeley
UNIVERSITY OF CALIFORNIA

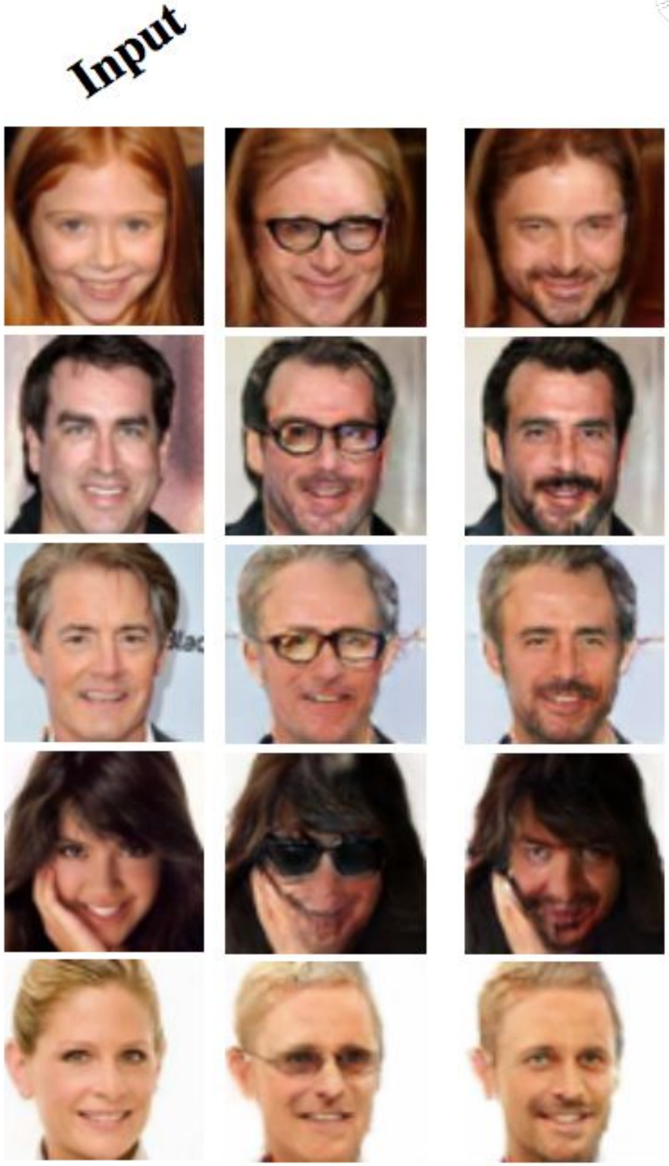


10
5
0
−5
−10
−10
−5
0
5
10
Second dimension of sampled latent variable mean z_2
First dimension of sampled latent variable z_1
9
8
7
6
5
4
3
2
1
0
Digit class

(a) Learned Frey Face manifold



(b) Learned MNIST manifold



Berkeley
UNIVERSITY OF CALIFORNIA

Variational Autoencoders (VAEs):

The deep learning perspective

Images from:

http://kvfrans.com/variational-autoencoders-explained/

Larsen, Anders Boesen Lindbo, et al. "Autoencoding beyond pixels using a learned similarity metric." arXiv preprint arXiv:1512.09300 (2015).

Manifold Hypothesis:

http://colah.github.io/posts/2014-10-Visualizing-MNIST/

Variational Autoencoders (VAEs):

The probability model perspective

Joint Probability, Bayes Rule

Variational Inference Approximation

How good is our approximation?

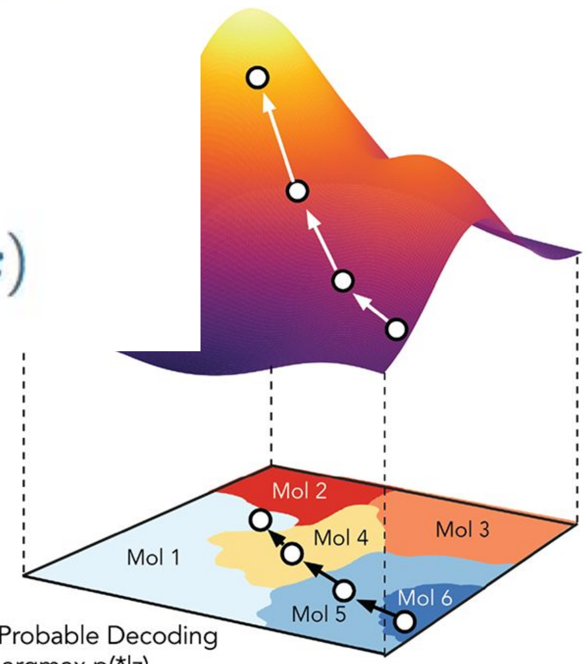
Kullback-Leibler divergence, ELBO, Jensen's inequality

Connection to neural net language

Gaussian Processes for Regression: aims

Goal: We know $(\mathbf{x}_i, y_i)$ for $i \in \{1, .., n\}$ and we want to estimate $y_\star$ for any $x_\star$, with $y_\star = f(x_\star) + \epsilon$.

Then we can find $x^{max} = argmax_x f(x)$



Most Probable Decoding
argmax p(*|z)

Gaussian Processes for Regression: model

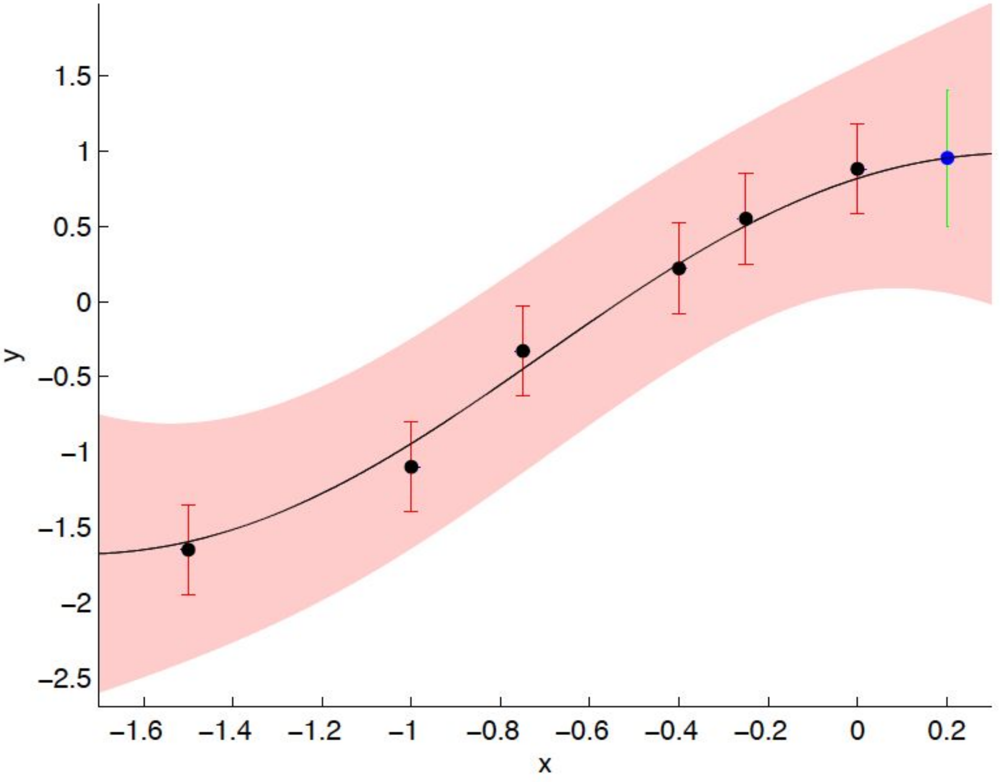
Model: $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, K)$
We assume a specific structure on K: $K_{(i,j)} = k(x_i, x_j)$,
with k a kernek function.
Then, by noting $K_\star = (k(x_1, x_\star), .., k(x_n, x_\star))$ and $K_{\star,\star} = k(x_\star, x_\star)$

We have $y_\star|\mathbf{y} \sim \mathcal{N}(K_\star K^{-1}\mathbf{y}, K_{\star,\star} - K_\star K^{-1} K_\star^T)$

Gaussian Processes for Regression: example

Example: $k(x, x') = \sigma_f^2 exp(\frac{-(x-x')^2}{2l^2})$
Closer points are more correlated.

Then you add the noise
$k_f(x, x') = +\sigma_n^2 \delta(x, x')$

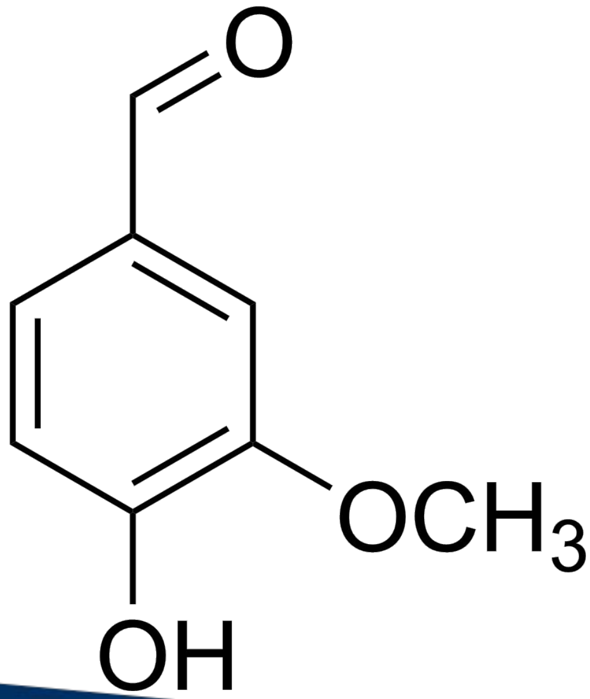


Berkeley
UNIVERSITY OF CALIFORNIA

SMILES: Simplified Molecular Input Line Entry Specification

- Switch from 2D to 1D structure unambiguously
- No hydrogen atoms

O=Cc1ccc(O)c(OC)c1



Autoencoder Framework

Results and Discussion

Conclusions and Future Works

Our Review/Comments?

Thank you for your attention!

09/10/2018

320 Soda Hall, University of California at Berkeley

CS294-150: Machine Learning and Statistics Meet Biology

Ryan Chung, Giulia Guidi, Weston Hughes, Hector Roux De Bezieux

Berkeley
UNIVERSITY OF CALIFORNIA

Sources

https://fr.wikipedia.org/wiki/Simplified_Molecular_Input_Line_Entry_Specification

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules; ACS Cent. Sci., 2018, 4 (2), pp 268–276

Tutorial - What is a variational autoencoder? https://jaan.io/what-is-variational-autoencoder-vae-tutorial/

Gaussian Processes for Regression:A Quick Introduction: https://arxiv.org/pdf/1505.02965.pdf

Tutorial on Variational Autoencoders https://arxiv.org/pdf/1606.05908.pdf

The original "Variational Autoencoder paper", https://arxiv.org/abs/1312.6114

Berkeley
UNIVERSITY OF CALIFORNIA

Supplementary slides

09/10/2018
320 Soda Hall, University of California at Berkeley

CS294-150: Machine Learning and Statistics Meet Biology

Ryan Chung, Giulia Guidi, Weston Hughes, Hector Roux de Bézieux

Berkeley
UNIVERSITY OF CALIFORNIA

More on Gaussian Processes for Regression

Example 2 and 3: We can have more complex kernels

$k_2(x, x') = \sigma_{f_1}^2 exp(\frac{-(x-x')^2}{2l_1^2}) + \sigma_{f_2}^2 exp(\frac{-(x-x')^2}{2l_2^2}) + \sigma_n^2 \delta(x, x')$

with $l_2 = 6l_1$

$k_3(x, x') = \sigma_f^2 exp(\frac{-(x-x')^2}{2l_1^2}) + exp(-2sin^2[\nu\pi(x-x')]) + \sigma_n^2 \delta(x, x')$