Text Generation using Variational Autoencoders (VAEs)

Aboubacry Mbargane, **Angela Musangi**
Elysee Manimpire, **Joseph Nnaemeka**
Junior Kaningini

Adviser: **Moustapha Cisse, Ph.D.**

African Institue for Mathematical Sciences

AIMS | African Institute for Mathematical Sciences
NEXT EINSTEIN INITIATIVE

April 21, 2022

Outline

1. Introduction
2. Method
3. Model Building
4. Conclusion

Project Objective

Construct a simple variational autoencoder (VAE) that will be used to convert the inputs and generate constrained samples that will be used to reproduce the optimal outputs closer to the inputs compared to the autoencoder.

Requirement for implementation

- **Dataset**: In this project, we used SMILES (Simple Molecules Inputs Line Entry System) dataset from chemical database.
- **Python Library**
- **Encoder Design**: Neural network for generating latent space to be sampled.
- **Decoder Design**: Neural network for reconstructing the target from samples of features.

Why Autoencoder?

Definition 1.1

Autoencoder: It is an essentially a neural network that is designed to learn an identity function in an unsupervised way such that it can compress and reconstruct an original input, and by doing that it discovers a more efficient and compressed representation of the original input data.

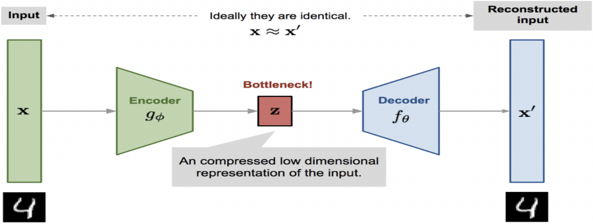

Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
$\mathbf{x}$
Encoder
g_ϕ
Bottleneck!
$\mathbf{z}$
Decoder
f_θ
$\mathbf{x}'$
An compressed low dimensional representation of the input.
4
4

Figure 1: Illustration of the Autoencoder architecture

Why Variational Autoencoder?

Definition 1.2

Variational Autoencoder(VAE): It is the same as autoencoder except that, the latent space of VAE is constrained such that it is continuous and allows easy random sampling and interpolation which supports also in backward propagation, which makes it better than autoencoder that has discontinuous latent space.

Variational Autoencoder

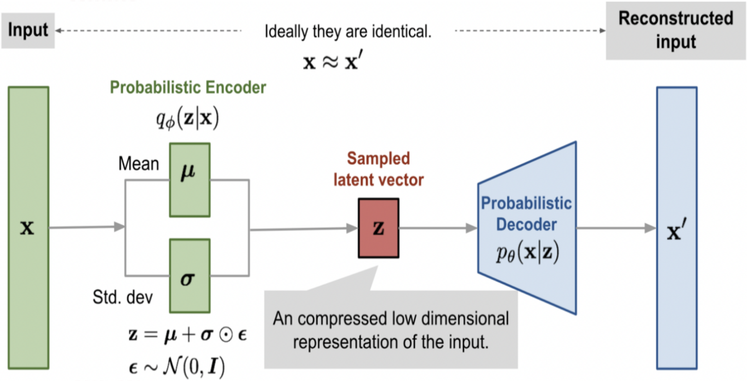

Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
Probabilistic Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
Mean
$\boldsymbol{\mu}$
$\mathbf{x}$
$\boldsymbol{\sigma}$
Std. dev
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
Sampled latent vector
$\mathbf{z}$
An compressed low dimensional representation of the input.
Probabilistic Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$

Figure 2: Illustration of the VAE model

Picture of task

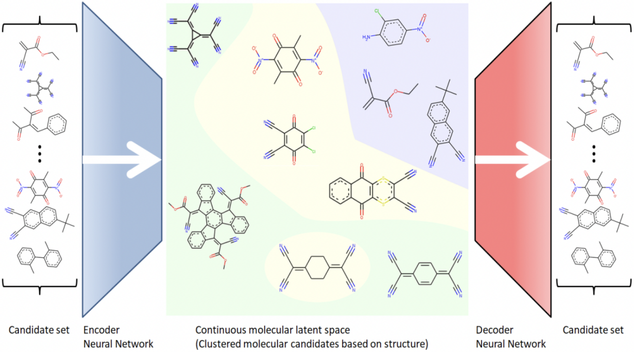
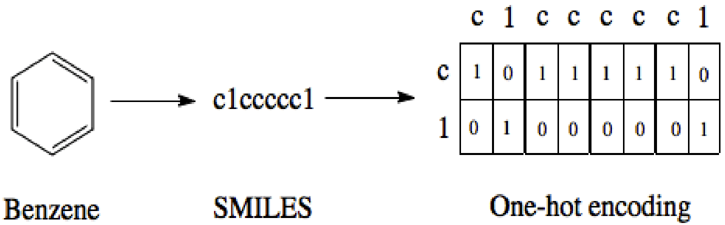

Candidate set
Encoder
Neural Network
Continuous molecular latent space
(Clustered molecular candidates based on structure)
Decoder
Neural Network
Candidate set

Figure 3: Illustration of the VAE model for SMILES dataset

Data Preprocessing

SMILES dataset was used to be encoded and generate new molecules. The structure of the chemical compounds were converted into the line of elements based on the shape, ring, atoms, and bonds. After by using one hot encoder, data have been transformed into numerical values. Example for Benzene encoding:



	c	1	c	c	c	c	c	1
c	1	0	1	1	1	1	1	0
1	0	1	0	0	0	0	0	1

Figure 4: One hot encoding of SMILES dataset

Model training

Model is trained to minimize the total loss which is made by summing up the expected difference between real and reconstructed data,negative log likelihood term with respect to encoder distribution and Kullback-Leibler divergence term between two probability distributions respectively.

$$\boxed{\mathcal{L}(\theta,\phi) = ||x-\hat{x}||_2^2 - \mathbb{E}_{z\sim Q_\phi(z|x)}[\log P_\theta(x|z)] + KL(Q_\phi(z|x)||P_\theta(z))}$$

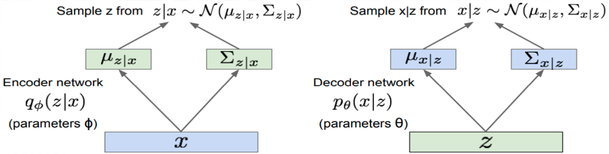


Figure 5: Illustration Distributions to be studied

Posterior Distribution

Given that we have the data x and unobserved (latent/hidden) space z.

- Assume z determine x.
- We have interest in the value of $P_\theta(z|x)$.
- True posterior mostly is intractable. We will use variational inference to estimate $P_\theta(z/x)$:

$$P_\theta(z|x) = \frac{P_\theta(x|z)P_\theta(z)}{P_\theta(x)} = \frac{P_\theta(x|z)P_\theta(z)}{\int_z P_\theta(x,z)dz}$$

- if z is continuous and many variables, the denominator might tend to infinity. This will make the network intractable, which is a special problem.

Posterior Distribution

Solution for intractability:

- Adding new tractable network represented by the distributional function $Q_{\phi}(z/x)$ to estimate the intractable distribution $P_{\theta}(z/x)$ such that $P_{\theta}(z|x) \approx Q_{\phi}(z)$.
- Now, we can sample z from $Q_{\phi}(z/x)$ since it can be tractable.

Similarity between distributions

Kullback Leiblier Divergence (KL): The measure of distance between two probability distributions.

- Given two discrete probability distributions P and Q on the same probability space X, then this distance will be represented as:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log P(x) - \sum_{x \in X} P(x) \log Q(x)$$

- From above, $D_{KL}(Q(z)||P(z|x)) = \mathbb{E}\big[\log Q(z)\big] - \mathbb{E}\big[\log P(z|x)\big]$.
- Since, $\mathbb{E}_{z \sim Q}\big[f(x)\big] = \sum_{x \in X} Q(x) f(x)$.

Variational inference

Now, by variational inference, we can estimate $P_\theta(z|x)$ as follow:

$$
\begin{aligned}
D_{KL}(Q_\phi(z|x)||P_\theta(z|x) &= \sum_z Q_\phi(z|x) \log \frac{Q_\phi(z|x)}{P_\theta(z|x)} \\
&= \mathbb{E}_{z\sim Q_\phi(z|x)}\Big[\log \frac{Q_\phi(z|x)}{P_\theta(z|x)}\Big] \\
&= \mathbb{E}_{z\sim Q_\phi(z|x)}\Big[\log Q_\phi(z|x) - \log P_\theta(z|x)\Big] \\
&= \mathbb{E}_{z\sim Q_\phi(z|x)}\Big[\log Q_\phi(z|x) - \log \frac{P_\theta(x|z)P_\theta(z)}{P_\theta(x)}\Big] \\
&= \mathbb{E}_{z\sim Q_\phi(z|x)}\Big[\log Q_\phi(z|x) - \log P_\theta(x|z) \\
&- \log P_\theta(z) + \log P_\theta(x)\Big]
\end{aligned}
$$

Variational inference

Let's $D_{QP} = D_{KL}(Q_\phi(z|x)||P_\theta(z|x))$

$$\begin{aligned} D_{QP} - \log P_\theta(x) &= \mathbb{E}_{z\sim Q_\phi(z|x)}\big[\log Q_\phi(z|x) - \log P_\theta(x|z) - \log P_\theta(z)\big] \\ &= -\mathbb{E}_{z\sim Q_\phi(z|x)}\big[\log P_\theta(x|z)\big] + \mathbb{E}_{z\sim Q_\phi(z|x)}\Big[\log \frac{Q_\phi(x|z)}{P_\theta(z)}\Big] \end{aligned}$$

Now, we get to the objective function of VAE, we can compute the loss, which is negative of objective function, $-(D_{QP} - \log P_\theta(x))$.

Kullback Leiblier divergence properties

- $KL(P||Q) \geq 0.$
- $KL(P||Q) \neq KL(Q||P)$

Variational Loss

- Loss:

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_z\big[\log P_\theta(x|z)\big] + D_{KL}(Q_\phi(z|x)||P_\theta(z)$$

- Loss=negative reconstruction likelihood + regularization loss.
- Maximizing the evidence lower bound, $(ELBOW(\theta, \phi))$ is the same as minimizing the loss.
- Minimization of loss, parameters have to be computed:

$$\theta^\star, \phi^\star = \arg\min_{\theta,\phi} \mathcal{L}(x^{(i)}, \theta, \phi)$$

Reparameterization Trick

- **Problem**: Discontinuous latent space due to the presence of stochastic node. Now, backpropagation is not possible.
- **Solution**:Create a continuous latent space by making a small modification to z.

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

- μ is the mean and σ is the standard deviation that have been generated by encoder.
- The epsilon, ϵ is the Gaussian noise that was sampled at zero mean vector and Identity covariance matrix from Gaussian distribution.

Pros of Variational Autoencoder (VAE)

- VAE is a powerful generative model due to the presence of probabilistic encoder. For instance, in the image generation.
- From the generation conditions, it is much easier working with VAEs.
- It is not surprising that the VAE always gets great likelihood, because they trained for a goal based on likelihood.

Cons of Variational Autoencoder (VAE)

- VAEs for text generation still has a lot of challenges because of the discrete nature of text data, and is a hot area of research.
- In image reconstruction, VAE produces the blurred images due to the direct computation of mean square error between real images and reconstructed images.

Conclusion

The variational autoencoder (VAE) is a framework for training two neural networks: an encoder and a decoder, in order to learn a mapping from a high-dimensional data representation to a lower-dimensional space and back again.

The lower dimensional space is called the latent space, which is often a continuous vector space with a normally distributed latent representation. The VAE parameters are optimized to encode and decode the data by minimizing the reconstruction loss while also minimizing a KL divergence term resulting from the variational approximation that can be interpreted as a regularization term.

References

1. Ali Ghodsi: *Deep Learning, Variational Autoencoder (Oct 12 2017)*
2. Stanford CS231n: *Lecture on Variational Autoencoders*
3. https://blog.bayeslabs.co/2019/06/04/All-you-need-to-know-about-Vae.html

No. :
Date :
Thank You For Your Attention
No. :