Lecture 19: Generative Models

Course: Biomedical Data Science

Parisa Rashidi

Fall 2019

Outline

- Overview
- Code

Credit

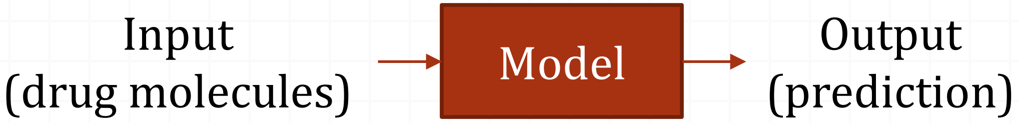
Slides are partially based on the material in Ramsundar, Bharath; Eastman, Peter; Walters, Patrick; Pande, Vijay. Deep Learning for the Life Sciences, Chapter 9.

Recent News

- https://www.eurekalert.org/pub_releases/2019-09/dka-abi090319.php

What are generative models?

- So far, our models:

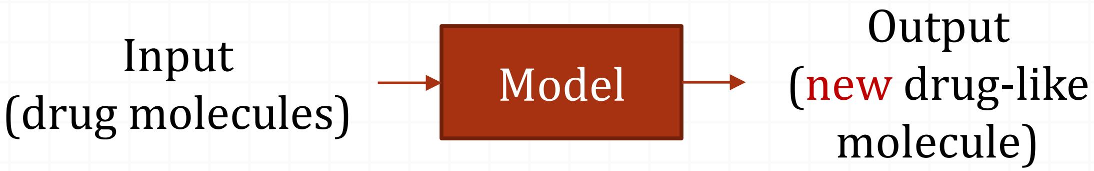


Input
(drug molecules)

Model

Output
(prediction)

- Generative models produce a sample as output.



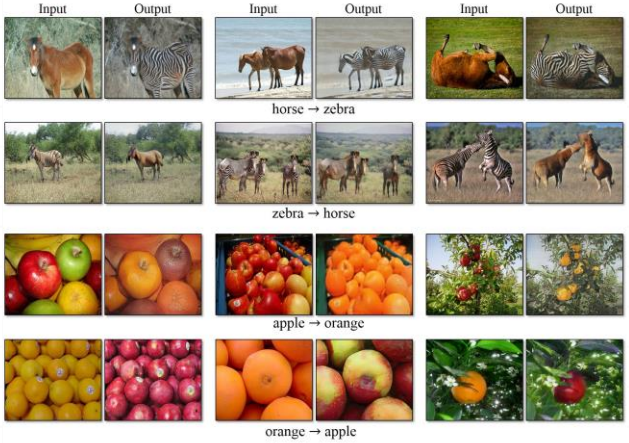
Input
(drug molecules)

Model

Output
(new drug-like molecule)

Generative Models

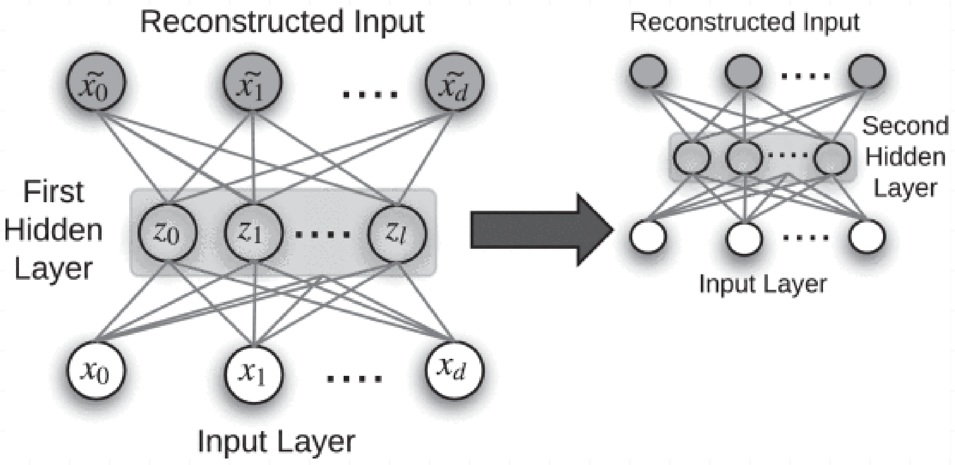
- These models are trained on samples drawn from a from some (possibly very complex, unknown) probability distribution.
- A generative model will produce new samples from that probability distribution.



Reed, Scott, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. "Generative adversarial text to image synthesis." arXiv preprint arXiv:1605.05396 (2016).

Autoencoder

- An autoencoder tries to make its input equal to the output.
- We adjust its parameters such that each input sample is as close as possible to the output sample.



Autoencoder

- The middle layer is usually called the latent space of the autoencoder.
- It is the space of compressed representation of samples (i.e. their codes).
- There is an encoder and an decoder.

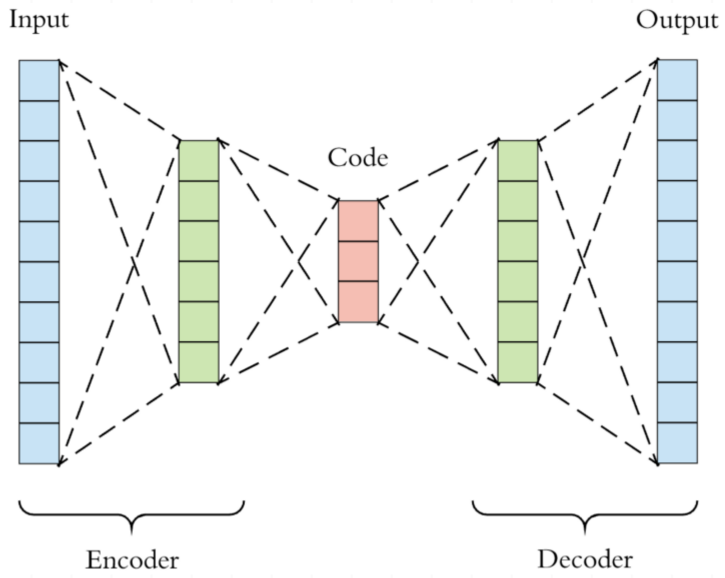
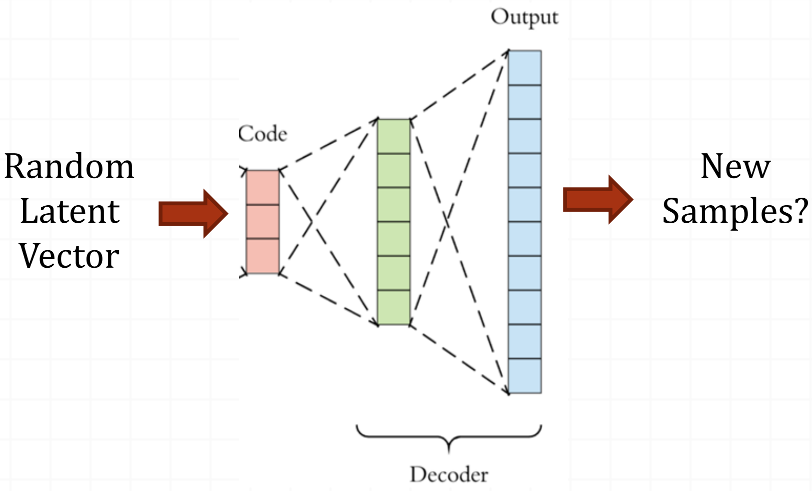


Image Credit: https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368

Generating New Samples with Autoencoders

- What if you were taking the decoder portion, and passed it random vectors in the latent space?
  - The encoder may only produce vectors in a small region of the latent space.



Random Latent Vector

New Samples?

Variational Autoencoder (VAE)

- A variational autoencoder (VAE) adds two features to address this problem.
  - It adds a term to the loss function to force the latent vector to follow a specific distribution (most often a standard Gaussian).
  - We add some random noise to the latent vectors to prevent the decoder from overfitting to details.

Generative Adversarial Networks (GANs)

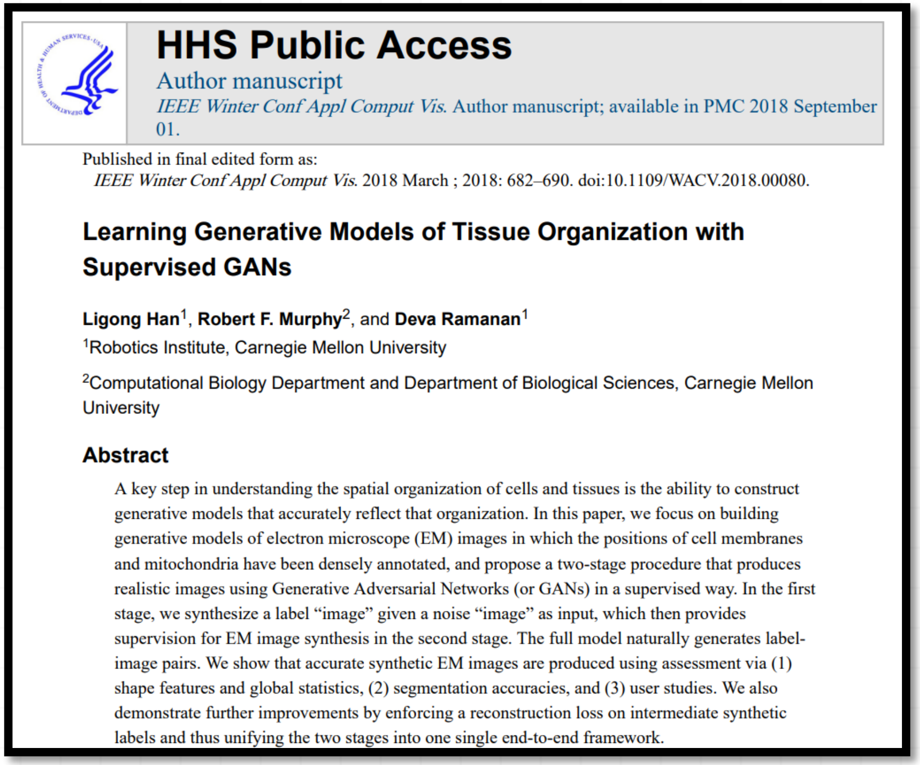
- It directly evaluates how well the generated outputs follow the expected distribution.
- A GAN consists of two parts: a generator and a discriminator.
- The generator takes random vectors and generates synthetic samples.
- The discriminator tries to distinguish the generated samples from real training samples, outputting the probability that it is real.

Training GANs

- Both parts are trained simultaneously.
- Generated (fake) samples
  - The parameters of the generator are adjusted to make the discriminator output as close as possible to 1.
  - The parameters of the discriminator are adjusted to make its output as close as possible to 0.
- Real data
  - The discriminator’s parameters are adjusted to make the output as close as possible to 1.
- It is almost as a competition between the two networks (i.e. adversarial)

Applications

- GANs and VAEs can be helpful in many application domains, e.g. drug, enzyme, or generally protein design.
- Unlike small molecules, it can be challenging for human experts to predict the downstream effects of changing a protein.



Published in final edited form as:

IEEE Winter Conf Appl Comput Vis. 2018 March ; 2018: 682–690. doi:10.1109/WACV.2018.00080.

Learning Generative Models of Tissue Organization with Supervised GANs

Ligong Han[1], **Robert F. Murphy**[2], and **Deva Ramanan**[1]

[1]Robotics Institute, Carnegie Mellon University

[2]Computational Biology Department and Department of Biological Sciences, Carnegie Mellon University

Abstract

A key step in understanding the spatial organization of cells and tissues is the ability to construct generative models that accurately reflect that organization. In this paper, we focus on building generative models of electron microscope (EM) images in which the positions of cell membranes and mitochondria have been densely annotated, and propose a two-stage procedure that produces realistic images using Generative Adversarial Networks (or GANs) in a supervised way. In the first stage, we synthesize a label "image" given a noise "image" as input, which then provides supervision for EM image synthesis in the second stage. The full model naturally generates label-image pairs. We show that accurate synthetic EM images are produced using assessment via (1) shape features and global statistics, (2) segmentation accuracies, and (3) user studies. We also demonstrate further improvements by enforcing a reconstruction loss on intermediate synthetic labels and thus unifying the two stages into one single end-to-end framework.

Generating New Molecules

Generating New Molecules

- We will generate new SMILES strings.
  - Pros: Simple to work with.
  - Cons: complex grammar, if the model does not learn all the subtleties, most strings will be invalid.

Dataset & Model

- We will use the MUV dataset,
  - 74,469 molecules of varying sizes.
- We will use a published model which uses a convolutional network for encoding and a recurrent network for decoding.

Generating New Samples

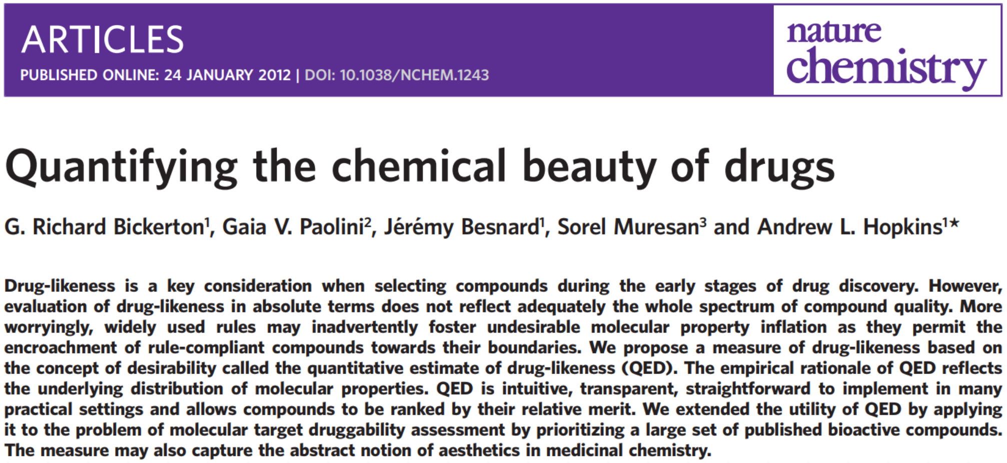
- See the colab notebook
- Note that many generated strings might not be valid molecules, we can check that using rdkit toolkit.
- Many valid molecules might not have drug characteristics.

Analyzing the Output

- One of the factors that we can check is size of the molecules,
  - Smaller than 10 atoms: unlikely to generate sufficient interaction energy to produce a measurable signal in the biological assay.
  - More than 50 atoms: might not be capable of dissolving in water.

Choosing among Generated Samples

- In practice, there are a number of ways to evaluate the quality of generated molecules.
- Quantitative Estimate of Drug-likeness (QED)
  - Scores molecules by comparing a set of properties and comparing the distributions to the same properties in marketed drugs.
- We can use RDKit to calculate QED.



ARTICLES

PUBLISHED ONLINE: 24 JANUARY 2012 | DOI: 10.1038/NCHEM.1243

nature chemistry

Quantifying the chemical beauty of drugs

G. Richard Bickerton[1], Gaia V. Paolini[2], Jérémy Besnard[1], Sorel Muresan[3] and Andrew L. Hopkins[1]*

Drug-likeness is a key consideration when selecting compounds during the early stages of drug discovery. However, evaluation of drug-likeness in absolute terms does not reflect adequately the whole spectrum of compound quality. More worryingly, widely used rules may inadvertently foster undesirable molecular property inflation as they permit the encroachment of rule-compliant compounds towards their boundaries. We propose a measure of drug-likeness based on the concept of desirability called the quantitative estimate of drug-likeness (QED). The empirical rationale of QED reflects the underlying distribution of molecular properties. QED is intuitive, transparent, straightforward to implement in many practical settings and allows compounds to be ranked by their relative merit. We extended the utility of QED by applying it to the problem of molecular target druggability assessment by prioritizing a large set of published bioactive compounds. The measure may also capture the abstract notion of aesthetics in medicinal chemistry.

More on Quality Check

- We can observe the fraction of molecules that obey the standard rules of chemical valance (e.g. carbon atoms have fur bonds, oxygen atoms have two bonds, ..)
- Still, they might contain a functional group that easily decomposes (e.g. hemiacetal).
- The synthesis of some more complex molecules might require more than 20 steps, still mostly driven by human intuition.
  - A few recent deep learning models learn the relation between product molecules and reaction steps.

Future

- Existing models are still challenging to work with and more robust models are needed.

- Nonetheless, in the future, we might be able to develop full generative models of tissues, embryonic development, modeling impacts of CRISPR, climate change, physics, etc.!

Ethical Challenges

- Let’s discuss some societal implications of generative models.