**Molecular Generation Using Convolutional Variational Autoencoders and Generative Adversarial Networks**

Contents

* Dataset
* Preprocessing
* Training
  + CVAE
  + GAN
  + Other Architectures
* Deployment
* Applications
* Research

Dataset

<https://www.nature.com/articles/s41597-022-01142-7>

Dataset consists of CSV file of molecules and their SMILES Representation, along with some other data.

Pre-Processing

* Rescales all molecules to standardized size
* Gets the width and height of all newly generated molecules
* Get 99th Percentile of these dimensions
* Standardizes the scale throughout images to fit in target size image
* Any in the top 1% of size are rotated through 45 Degrees as an approximation fix

Training

Hyper parameters – 24 EPOCHS, 48,000 Training Images, 64 Batch Size

* Convolutional Variational Auto Encoder
* Generative Adversarial Network

Deployment

Applications

Research

Process

* Preprocess RDKit Generated Images – Remove Glitched Images
* Encode Smiles Using Language Model/Vector/Bert
* Use Smiles Vector as Input to Variational Autoencoder, x\_train
* Train it off Using RDKit Generated Images as y\_train
* Different Perspectives
* Remove smiles truncating names
* Remove Molecules that Do not Generate Properly

Diffusion Models

Smile + Other Information (As Vector) -> Skeleton (Using VAE) (loss from RDKit Images)

Skeleton Image to Smile -> Trained off RDKit Images

Trained

Model 1 – Smile to Image Representation

Model 2 – Image to Smile Representation

Goal, Input a smile with some parameters like make it long and user is generated image of skeletons

Pipeline

SMILES Representation -> ChemBERTa or RDKit Pretrained Embedding -> Vector

Vector -> Auto-Encoder -> Skeleton (target is RDKit Generated Version) -> Conv Net -> SMILEs

<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0286-7>

Zinc Dataset

<https://arxiv.org/pdf/2209.01712.pdf>

CUDA – for optimisation

Vectors not lining up with right file

Bugs -> Skeletons not generated well

T-Sne or PCA

PCA is way quicker to remove dimensions but slower

Add In Elbow Method For Visualization and clustering

KMeans

DBScan

MeanShift

[OPTICS](https://scikit-learn.org/stable/modules/clustering.html#optics)

Bisecting-KMeans

AUTOMATIC GRAPH SAVING

Visualising

Dimensionality Reduction

* PCA
* TSNE

Clustering

* Kmeans
* Agglomerative

Full Process

Preprocessing

Training

Generation

Multiple models

* Purely Smile to Image generation
* Smile With Conditions Generation

Other Visualisation Techniques

Prep For Meeting

* Visualisations
* Work out Baskerville
* Generate some test images
* Begin work on academic paper
* Work Out Something I Want to do with the model

Components

* Dataset Creation
* Preprocessing Images
* Get Conditions
* Train Model
* Generate Images
* Academic Paper
* Refactor Code
* Turn Into Full Pipeline

To Improve

* Use bigger embedding model
* Use coloured images
* Train for longer
* More detailed architecture
* Larger Batches
* More Training Data

Generation Options

* Generate New Random Molecule
* Generate New Molecule from starting molecule and Condition
* Generate New Molecule from noise and condition

Code Review

* Error Handling
* Commenting
* Documentation
* Efficiency Checks
* Progress Bars
* UI
* Remove Unnecessary Code
* Organise import’s

Questions

* For new molecule generation, is it better to concat starting molecule with its own conditions or new conditions and then pass through model concating new conditions with latent dimensions or new conditions from the start

Main.py should have 1 function for each component

* Generate Dataset
* Generate NLP Model for Input synthesis from natural language
* Preprocess
* Train
* Post Process
* NLP model to generate starting molecules and conditions
* Deploy to users
* Generate multiple molecules from multiple starting molecules, describe the effects of the changes to the molecules

Turn It into a website with Django?

Diffusion models

Discord Bot

Use a language model to create the conditions

Goal Usage:

E.g 1

Input : Generate me a molecule starting from benzene that can be used to treat schizophrenia

Output : Here is 5 possible molecules that can be used to treat schizophrenia starting from a benzene ring

E.g 2

Input : Generate me some possible molecules for treatment of depression

A prompt has two key parts

Target of molecule (What its treating) and starting molecules (optional)

Before Meeting

* More Visualizations
* Talk Plan
* Learning Slurm/Baskerville
* Train 2 Models, one small one for showing the VAE generates clear Images
* Train a second on the full dataset overnight to show it can learn
* Generate Example Images

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570547/>

<https://www.kaggle.com/datasets/jithinanievarghese/drugs-related-to-common-treatments>?

SPLIT DATA INTO VALIDATION, TRAINING, TEST

List Target Illnesses

Get all currently patented treatments

Prompt

Design me 20 possible molecules for treating Parkinson’s with a faster onset then X drug and less biotoxic, using examples of

More Visualizations of Input Embedding and Latent Space

Make it robust so it works with any dataset

Small Dataset, both 44 and 194 is optimal number of clusters?

To Do

* Build Dataset Builder
* Visualisations
* Build Custom dataset
* Write Academic Paper
* Input Synthesis
* Images to Smiles
* Full Automatic Pipeline
* UI
* Do Cluster Evaluation of Large Dataset
* QSAR
* Protein Folding
* Create Full Drug Discovery Pipeline
* RL