De Novo Drug Design Using LSTM and Autoencoder

Xikun Liu xikun@ucsb.edu

Introduction:

De novo drug design means generating new drugs that have desired properties. The synthetically accessible and druglike chemical space is estimated to cover more than 10^{30} molecules and it's hard to fully explore with exiting molecules library.[1] Computational de novo drug design can help to explore the chemical space and generate new molecules with similar properties.

In this study, I compared autoencoder and long short-term memory (LSTM) in generating valid simplified molecular-input line-entry system (SMILES) strings. Autoencoder is an artificial neural network that composed of encoder and decoder. Encoder and decoder are usually a one layer or multiple layers neural network. The encoder converts the input data into a vector representation with important information, but reduced dimension and the decoder tries to reconstruct the input data. LSTM is a recurrent neural network that composed of input gate, forget gate and output gate. Input gate decides how much input can be feed into the hidden layer. Forget gate decides how much memory (previous input) will be added to the current input. Output gate decides how much hidden layer information will be delivered out of the cell. LSTM is able to produce different output with the same input because it has memory of the previous input. It's been used in natural language processing and also used to generate SMILES strings. Gupta et.al used LSTM to generate valid SMILES string with 97% success rate.[1]

Data：

A small molecular database with 1128 molecules and their water solubility called MolecularNet ESOL is used for both LSTM and autoencoder model. Each molecule is represented by a SMILES string and each character of the string is either represented by one unique integer or one-hot encoding. For labelling one unique integer, there are 100 characters in total and each character is represented by integer between 0 to 99. For one hot encoding, the database contains SMILES strings of a length of 1-97 character with 31 unique characters. 'E' is added at the end of a SMILE string to indicate the end of a string, which makes 32 unique characters and a maximum length of 98. Therefore, one SMILE string is represented by a matrix of 98*32 dimension with 0s as paddings.

Network:

Autoencoder has one fully connected linear layer as encoder and the same for decoder with ReLU as activation function. The input matrix of 98*32 dimension is mapped into a vector of 100 *1 and then decoded back to 98*32 dimension. The maximum value of each row is assigned to 1 and the rest are set to 0. Later it's translated back to SMILEs string by one-hot encoding. The character after 'E' or '##' was truncated to generate the final SMILE string. MSELoss is used as loss function and SGD is used as optimizer. The whole model is trained with 908 training data with 2,000 epochs and tested with 220 testing data.

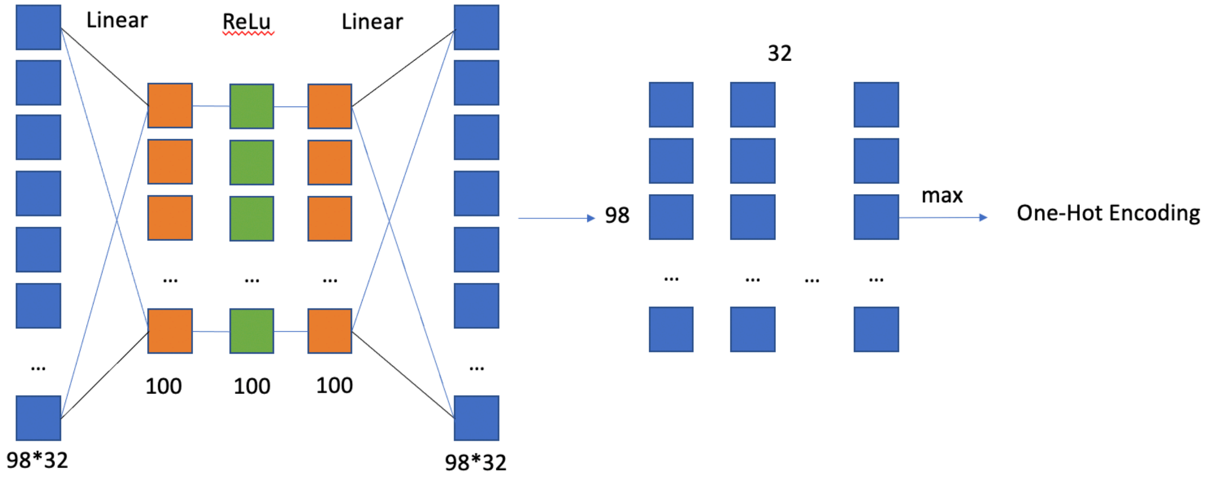


Figure 1: The structure for Autoencoder model

LSTM network is built similarly to Yasonik[2] but rather than 3 LSTM layers, I have 2 LSTM layers each of size 256. A fully connected layer was applied after the LSTM cells followed by a Softmax function with temperature parameter to calculate the probability. A multinomial probability distribution was used to select the predicted character. In this way, the character with the biggest number has the highest chance of being selected while allowing the model to try other less likely character. 250-character chunk of smiles text was randomly sampled with a batch size of 1. Backpropagation through time was used to train the network with the cross-entropy loss function and ADAM optimizer.

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y})$$

$$P(y_i) = \frac{\exp(y_i / T)}{\sum_{j=1}^{n} \exp(y_j / T)}$$

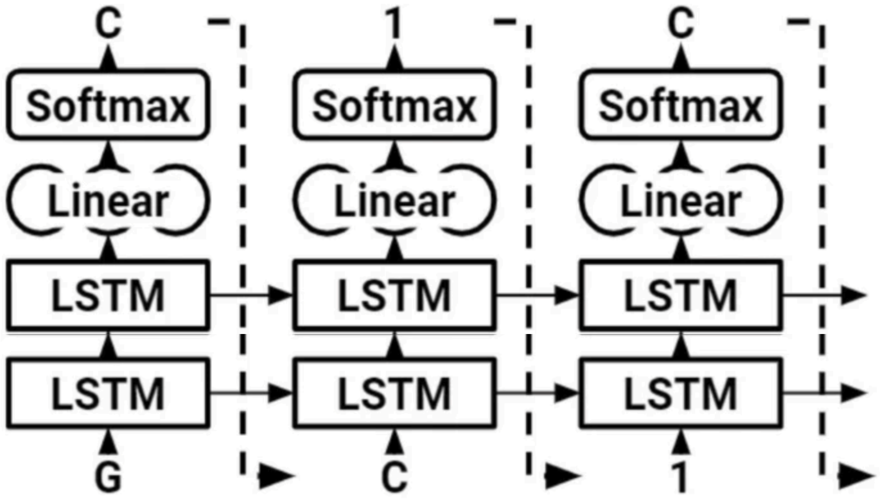


Figure 2. The loss function, softmax function with temperature parameter and the structure of the LSTM model used

1. Gupta, A.; Muller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G., Generative Recurrent Networks for De Novo Drug Design. *Mol Inform* **2018,** *37* (1-2).
2. McClenaghan, C.; Hanson, A.; Lee, S. J.; Nichols, C. G., Coronavirus Proteins as Ion Channels: Current and Potential Research. *Front Immunol* **2020,** *11*, 573339.

Training:

The autoencoder model was trained with 2000 epochs with a batch size of 64. The training loss decreased and plateaued after 2000 epochs and the average time per epoch is 234.7 ms.

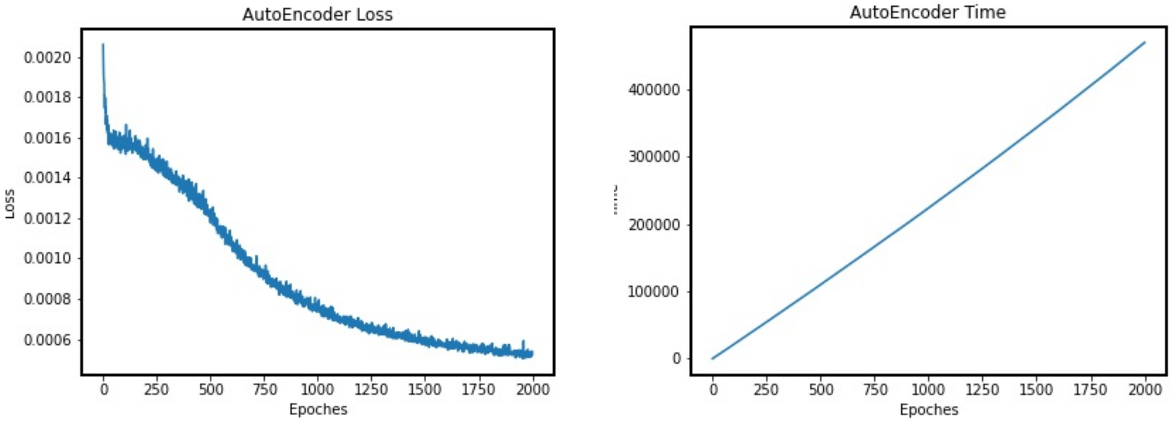


Figure 3. Loss function and training time for Autoencoder model with 2000 epochs

The LSTM model was trained with 5000 epochs with a batch size of 1. The training loss decreased with an average time of 237.4 ms per epoch.

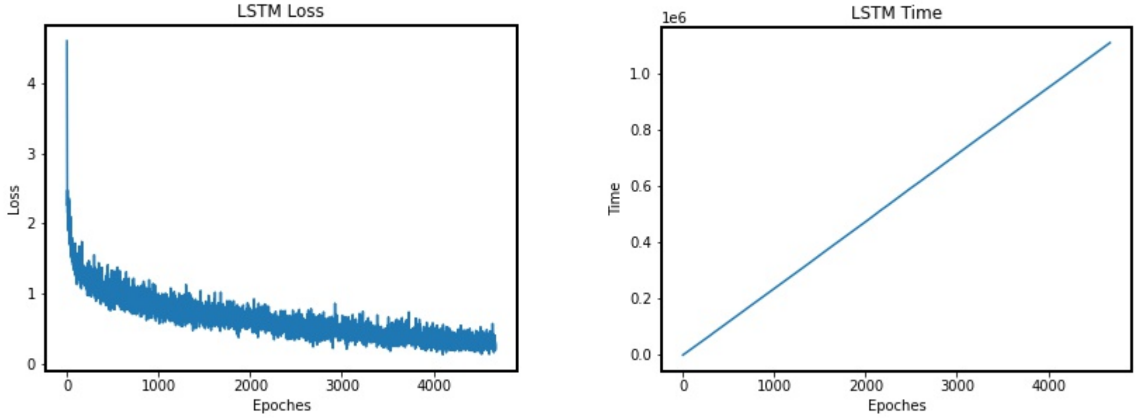


Figure 4. Loss function and training time for LSTM model with 5000 epochs

Validation:

Autoencoder:

After 2000 epochs, this model predicted 47 out of 902 molecules correctly with an accuracy of 5.21, and 88 out of 902 molecules are valid SMILES for training data.
For testing data, this model predicted 10 out of 226 molecules correctly with an accuracy of 4.42 and 27 molecules are valid SMILES. The result is very bad. It might be because there are too many zeros in the one hot encoding and it’s hard to train the model. The results don’t improve much after 2,000 epochs.
Besides, the correct predictions for SMILES are relatively simple with mostly carbons and oxygens.

After removing the encoder part, 900 1*100 vectors of random number between 0 to 1 was feed into the decoder to generate 900 molecules, 10% out of which are valid SMILES and the structure is relatively simple.

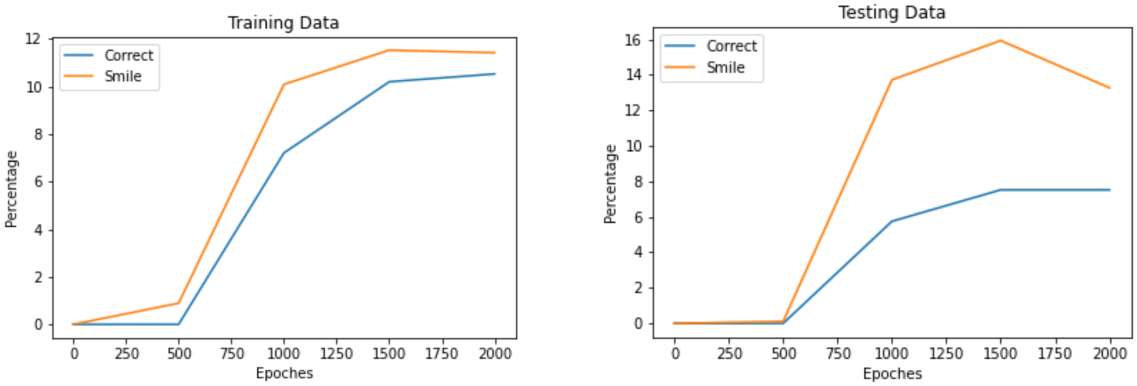


Figure 5 Rate of generating output same as input and rate of generating valid SMILES for training and testing data.

Table 1. Example of correctly predicted SMILES, valid SMILES and generated SMILES

Correctly Predicted SMILES	Valid SMILES	Generated SMILES
CCOC=O	CC#C	CO
COC(C)(C)C	C	CCCC
CCCCC(C)O	C=C1C#CCC1	C=CC
CCC(C)CCO	Nc1c#cc(=O)c1	C
CCOCC	OC	CC#C

LSTM

After 5000 epochs, LSTM achieves a correct rate of 94% for training data and 67 % for testing data. The correct rate increases sharply after 1000 epochs but decreases slightly after 3000 epochs for testing set indicating over-fitting. However, the change of generating legit SMILES string is higher after 5000 epochs.

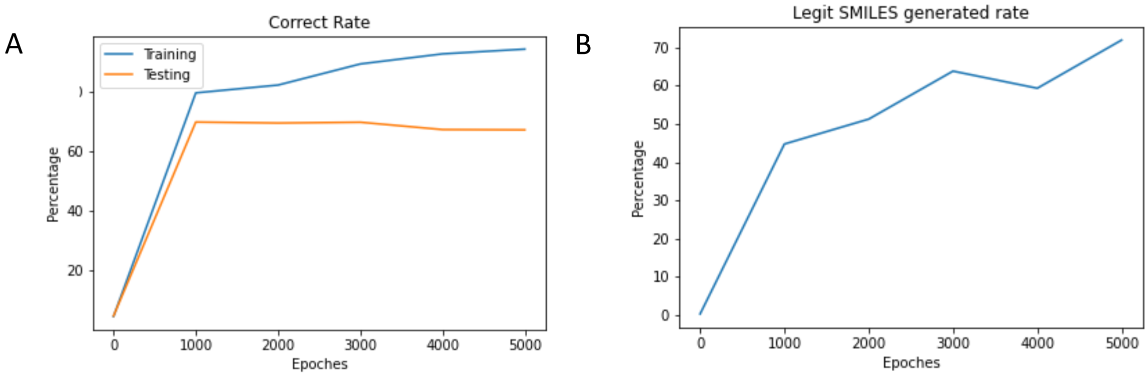


Figure 6. Rate of generating correct next character for training and testing data and the rate of generating valid SMILES after 5000 epochs

Out of 618 strings generated, the model shows a 72% of chance of generating legit SMILES and within the legit SMILES strings, 32% are duplicate and 40.6 % overlap with the training data. 172 new SMILES were generated with a rate of 28%.

To investigate the similarity between the training SMILES and generated SMILES, principal component analysis is carried out to visualize the SMILES in two dimensions. As we can see in Fig, the generated data overlap well with training data, which means the structure of the molecules generated are similar to the molecules used for training.

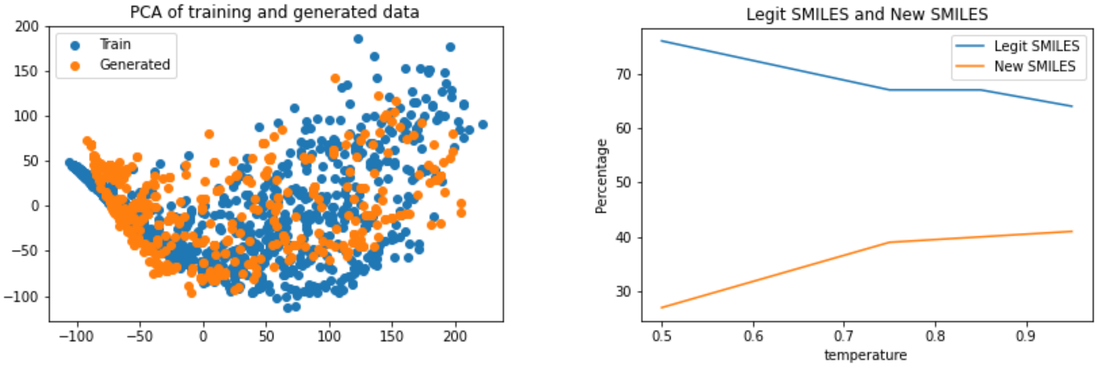
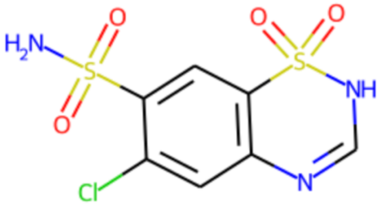
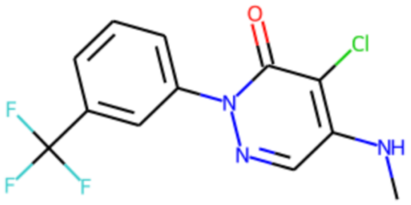


Figure 7A Training and generated data plotted in reduced dimensions by principal component analysis B The rate of generating valid SMILES and new SMILES under different temperature

The molecules generated using LSTM are also more complicated compared to Autoencoder. Some selected samples are plotted below:



NS(=O)(=O)c2cc1c(N=CNS1(=O)=O)cc2Cl



CNc2cnn(c1cccc(c1)C(F)(F)F)c(=O)c2Cl

Discussion:

LSTM with labelling performs much better than Autoencoder with one-hot encoding, achieving a chance of 72% of generating valid SMILES string compared to 10%. It might be LSTM performs better in handling text, or the feature used for Autoencoder is not appropriate. I will try variational autoencoder in the future since that’s a generative model rather than reproduce the input data.

The temperature used for LSTM was set to be 0.85. When the temperature increases, the model is more likely to try new character which increases the structure diversity but lower the chance of generating legit SMILES. When the temperature decreases, the model is more likely to generate legit SMILES but tend to play safe.