ENSC 813

Assignment 5: Project Update

Brett Hannigan
bchannig@sfu.ca

2019-03-01

1 Problem Definition

Two important problems in the field of synthetic chemistry are reaction prediction and its inverse, retrosynthesis. The reaction prediction problem involves predicting the most likely product of a reaction given a set of reactants. This is may be done using advanced chemical modelling tools that are computationally expensive to run. A deep learning system that is able to accurately predict products could be useful in predicting a large number of reactions in pharmaceuticals, process development, or research. In the retrosynthesis problem the series of reactants required to obtain the given product are selected in a desirable way, e.g. fewer steps, higher yields, and lower cost. Software packages that are used for this problem are generally designed using expert systems. Deep learning methods have recently become the state-of-the-art for retrosynthesis [1].

2 Background Research

Table 1 lists some literature related to this project.

Table 1: Some recent examples of deep learning applied to chemistry

Ref.	Year	Application	Dataset	Input Type	Architecture
[2]	2016	Prediction	Custom	Chemical fingerprints	Perceptron
[3]	2017	Retrosynthesis	Patents [4]	SMILES strings	seq2seq
[5]	2018	Autoencoder	QM9, ZINC database	SMILES strings	RNN
[6]	2018	Retrosynthesis	Reaxys database	Graphs	GCN

Two challenges are in picking the type of input and network architecture. Once common way of representing chemical compounds in a computer are SMILES strings, textual strings that encode the structure of a molecule. SMILES stings are easy to work with but may not convey the positional relationship between atoms as other methods. They also require correct grammar to be interpretable, thus many networks that operate on them require a final step of discarding invalid SMILES. Another approach is to use molecular fingerprints. In this method, a graph search algorithm records functional neighbourhoods of the molecule as hash values. Fingerprints more obviously reflect functional groups but the hashing function isn't invertible to obtain the exact structure from the fingerprint. The most natural and promising input type would be a relational graph, representing atoms as nodes and edges as bonds.

3 Requirements

Due to time constraints of the project and the difficulty of the retrosynthesis problem, I intend to begin by designing a molecular autoencoder using graph convolutional networks (GCNs). This will allow experimenting with more advanced types of neural networks while keeping the problem reasonable in scope. If time permits, another network stage may be designed to perform reaction prediction or retrosynthesis taking as input tensors from the latent space encoder representation of the molecules.

4 Solutions and Development Work

So far, I have downloaded several of the datasets from the works in Table 1 and installed the cheminfomatics tool `RDKit`. Using this tool, I have written data preprocessing code that can convert between SMILES and graphs as well as output properties about the structure of a molecule for use as features. The chosen features implementation for nodes may include features such as [7]:

- Atomic number (one-hot vector for 1, 6–9, 15–17, 35, 53, or metal)
- Chirality (one-hot vector for none, R, or S)
- Formal charge
- Partial charge
- Ring size (none, 3–8)
- Hybridization (one-hot vector for sp^3, sp^2, or sp)
- Hydrogen bonding (one-hot vector for donor, none, acceptor)
- Aromaticity (binary)

Because graph convolutional layers are not provided in `Keras`, I have investigated options on how to define them. There are several implementations directly in `TensorFlow`, as well as `Keras` custom layers. The code for the latter, provided at `https://github.com/tkipf/relational-gcn`, seems like an appropriate starting point with the following issues:

- The network is designed to process a single large relational graphs (e.g. from a social network) rather than multiple small molecule graphs
- There are currently no feature vectors for edges
- There is no graph pooling layer

The next steps are to finish modifying this code and test it with a subset of the data. I will have to do more research into how operations such as convolution and pooling are done on graphs (e.g. [8], `https://towardsdatascience.com/how-to-do-deep-learning-on-graphs-with-graph-convolutional-networks-7d2250723780`). An updated timeline is shown in Table 2

Table 2: Revised timeline

Date	Milestone
2019-01-31	Define the problem
2019-02-09	Do background research
2019-02-12	Specify requirements
2019-02-21	Brainstorm solutions
2019-02-28	Choose the best solution
2019-03-15	Do development work
2019-03-22	Build a prototype
2019-04-06	Test and redesign

References

[1] M. H. Segler and M. P. Waller, "Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction," *Chemistry - A European Journal*, vol. 23, no. 25, pp. 5966–5971, 2017.

[2] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, "Neural networks for the prediction of organic chemistry reactions," *ACS central science*, vol. 2, no. 10, pp. 725–732, 2016.

[3] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, and V. Pande, "Retrosynthetic reaction prediction using neural sequence-to-sequence models," *ACS central science*, vol. 3, no. 10, pp. 1103–1113, 2017.

[4] D. Lowe, "Chemical reactions from US patents (1976-Sep2016)," 2017.

[5] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[6] M. H. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI," *Nature*, vol. 555, no. 7698, pp. 604–610, 2018.

[7] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of Computer-Aided Molecular Design*, vol. 30, no. 8, pp. 595–608, 2016.

[8] W. B. W. Vos, "End-to-end learning of latent edge weights for Graph Convolutional Networks," 2017.