Intro to Feature Representation in Virtual Screening

Shengchao Liu, Gitter Group

Feature Representation

1. Raw Molecule Representation (Graph CNN)
   a. atom info
   b. bond info
2. SMILES (RNN, CNN, RNN+CNN)
   a. string, a sequence of characters
   b. lack of structural info
3. Morgan Fingerprints/ECFP (Dense NN, classical ML)
   a. hashing will lose more information

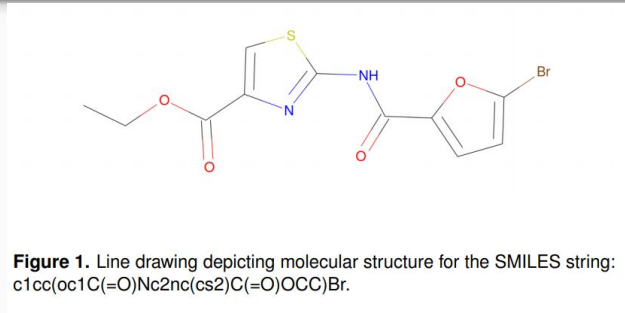
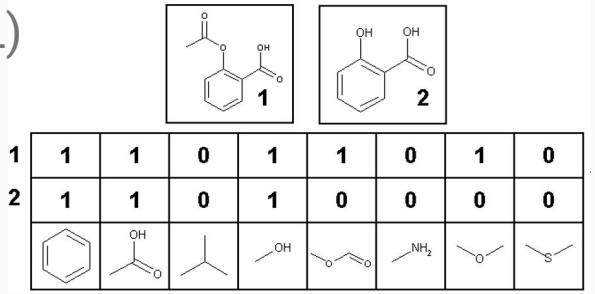
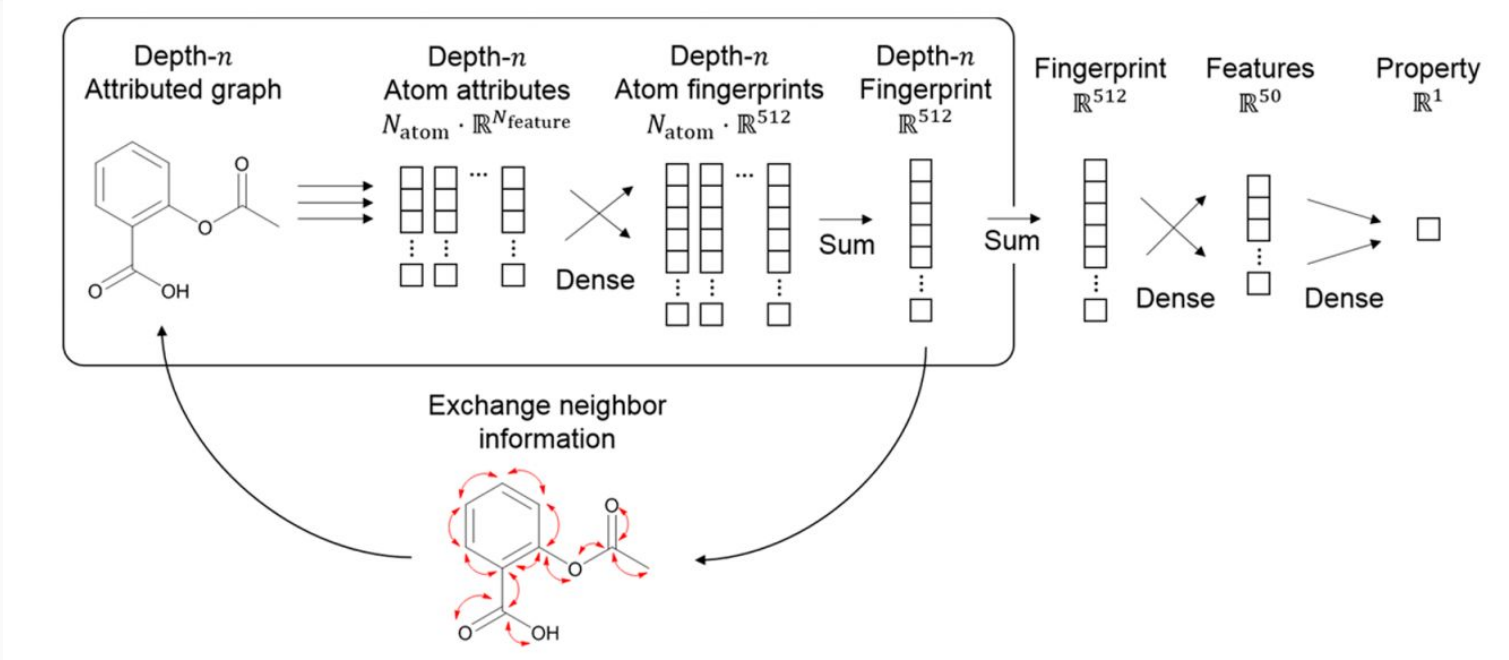


Figure 1. Line drawing depicting molecular structure for the SMILES string: c1cc(oc1C(=O)Nc2nc(cs2)C(=O)OCC)Br.



1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0

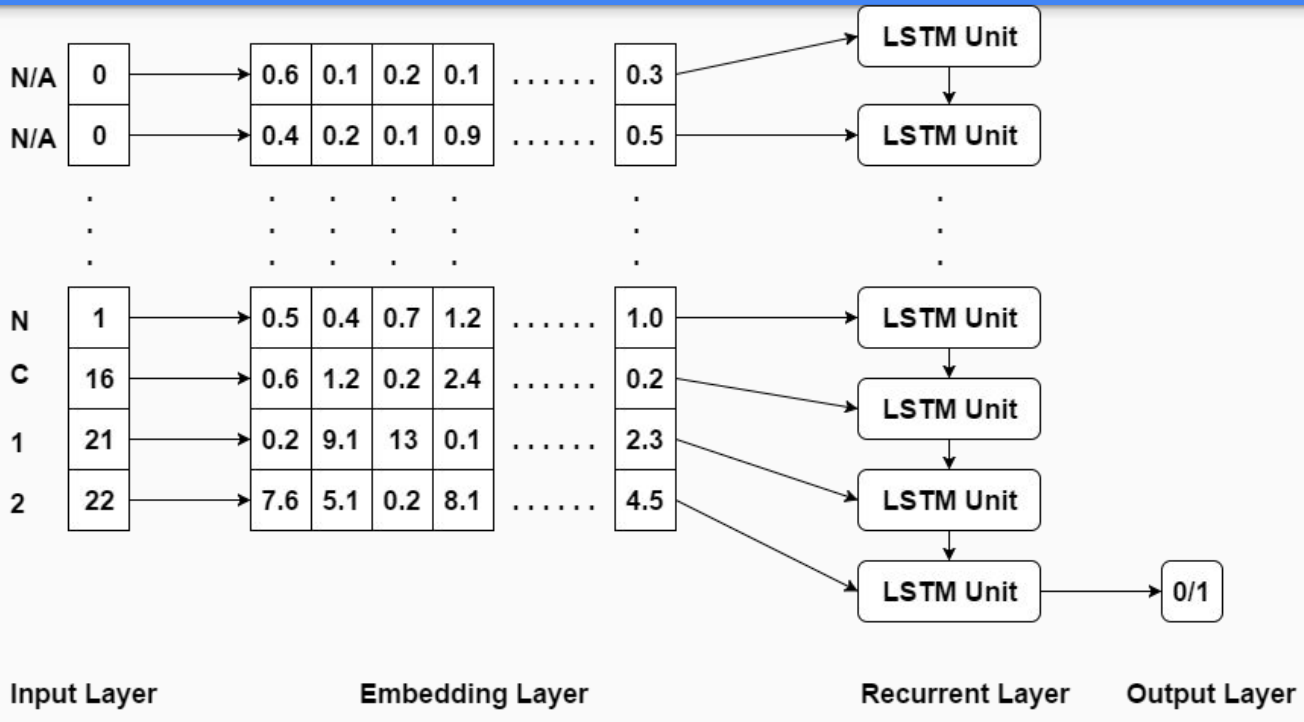
Graph CNN Framework



Related Work

- Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." Advances in neural information processing systems. 2015.
- Niepert, Mathias, Mohamed Ahmed, and Konstantin Kutzkov. "Learning convolutional neural networks for graphs." International Conference on Machine Learning. 2016.
- Kearnes, Steven, et al. "Molecular graph convolutions: moving beyond fingerprints." Journal of computer-aided molecular design 30.8 (2016): 595-608.
- Coley, Connor W., et al. "Convolutional embedding of attributed molecular graphs for physical property prediction." Journal of chemical information and modeling 57.8 (2017): 1757-1772.

RNN Framework



Related Work

- Jastrzębski, Stanisław, Damian Leśniak, and Wojciech Marian Czarnecki. "Learning to SMILE (S)." arXiv preprint arXiv:1602.06289 (2016).
- Jaeger, Sabrina, Simone Fulle, and Samo Turk. "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition." Journal of chemical information and modeling(2017).
- Vanilla LSTM, (Keck Paper)

More oftenly used in molecule generation tasks, like GAN and AE.