Communicative Representation Learning on Attributed Molecular Graphs

Ying Song[1,2†], **Shuangjia Zheng**[1†], **Zhangming Niu**[2], **Zhang-Hua Fu**[3,4], **Yutong Lu**[1] and **Yuedong Yang**[1*]
[1]Sun Yat-sen University
[2]Aladdin Healthcare Technologies Ltd
[3]The Chinese University of Hong Kong, Shenzhen
[4]Shenzhen Institute of Artificial Intelligence and Robotics for Society
{songy75, zhengshj9}@mail2.sysu.edu.cn, zhangming@aladdinid.com, fuzhanghua@cuhk.edu.cn, yutong.lu@nscc-gz.cn, yangyd25@mail.sysu.edu.cn

Abstract

Constructing proper representations of molecules lies at the core of numerous tasks such as molecular property prediction and drug design. Graph neural networks, especially message passing neural network (MPNN) and its variants, have recently made remarkable achievements in molecular graph modeling. Albeit powerful, the one-sided focuses on atom (node) or bond (edge) information of existing MPNN methods lead to the insufficient representations of the attributed molecular graphs. Herein, we propose a Communicative Message Passing Neural Network (CMPNN) to improve the molecular embedding by strengthening the message interactions between nodes and edges through a communicative kernel. In addition, the message generation process is enriched by introducing a new message booster module. Extensive experiments demonstrated that the proposed model obtained superior performances against state-of-the-art baselines on six chemical property datasets. Further visualization also showed better representation capacity of our model.

1 Introduction

Accurately predicting the properties of molecules has always been a topic of interest in the pharmaceutical community. The major goal of molecular property prediction is to remove compounds which are more likely to have property liabilities during downstream development, hence the desire to save tons of resources as well as time [Cherkasov *et al.*2014].

Briefly, the key idea of property prediction is to first map an input molecule m to a dense feature vector with a representation function, $h = g(m)$, and then make prediction of the targeted property based on the embedding by $y = f(h)$.

Early studies of quantitative structure-property relationships (QSPR) have been carried out based upon feature engineering e.g. expert-crafted physicochemical descriptors [Nettles *et al.*2007] and molecular fingerprints [Rogers and Hahn2010]. However, descriptor-based representation methods presume that all information related to the task predictions is covered in the chosen descriptor set, restricting the capability for a model to learn beyond the existing chemical knowledge.

In recent decades, with the substantial increase in available experimental molecular properties data points, machine learning especially deep learning methods have shown strong potentials to compete with or even outperform conventional approaches. Compared to the previous descriptor-based methods, deep learning-based models can take the relatively lossless 'raw' molecule formats e.g. SMILES strings and topological graphs as input, and then train models in an end-to-end fashion to predict the target endpoints. The representations obtained from these models are potentially able to profile comprehensive information for molecules.

A chemical structure could be intrinsically depicted as a hydrogen-depleted topological graphs whose nodes represent atoms with edges representing for bonds. In this sense, graph-based algorithms could be intuitively introduced to learn the representations of molecules. [Duvenaud *et al.*2015] reported a neural fingerprint method as an alternative to molecular fingerprints, and also one of the earliest efforts in employing graph convolution approaches on chemical representations. Then, several graph convolution models were reported as extensions to the Duvenaud's method by increasing molecular attributes. Recently [Gilmer *et al.*2017] summarized a general architecture called message passing neural networks (MPNNs) that demonstrated superior performance in predictions of quantum chemical properties. Broadly speaking, the MPNN framework includes three main modules: (1) message passing module, where, information of each atom is transmitted from its neighbors across the molecular graph into a message vector; (2) updating module, where the hidden states at each atom in the molecule are updated based on the obtained message vector;

However, MPNN and its variants mainly focused on obtaining effective vertices (atoms) embedding, but ignored the information carried by edges (bonds) that can be favorable to many downstream tasks such as node or edge embeddings and graph representations. To alleviate this problem, directed MPNN (D-MPNN) [Yang *et al.*2019] has been introduced to alleviate the problem by using messages associated with

[†]These two authors contributed equally.
[*]Corresponding author.