Revised title and Abstract (05.21.19)

Yingyu Liang
University of Wisconsin

"Title: N-Gram Graph: A Simple Unsupervised Representation for Molecules."

Abstract: Machine learning techniques have recently been adopted in various applications in medicine, biology, chemistry, and material engineering. An important task is to predict the properties of molecules, which serves as the main subroutine in many downstream applications such as virtual screening and drug design. Despite the increasing interest, the key challenge of constructing proper representations of molecules for learning algorithms remains largely unsolved. This paper introduces the N-gram graph, a simple unsupervised representation for molecules which benefits from recent embedding methods while preserving the simplicity of fingerprints. The method first embeds the vertices in the molecule graph. It then constructs compact representations for the graph by assembling the vertex embeddings in short walks in the graph, which we show is equivalent to a simple graph neural network that needs no training. The representations can thus be efficiently computed and then used with supervised learning methods for prediction. Experiments on 63 tasks on 10 benchmark datasets demonstrate its advantages over both popular graph neural networks and traditional fingerprint methods. This is complemented by theoretical analysis showing its strong representation power.