Semi-Supervised Deep Generative Model for Toxicology

Name: Tony (Tongzhou) Shen

SFU Faculty/Major: School of Computing Science, Molecular Biology and Biochemistry and Computing Science

Presentation Description:

Generate chemical molecules with salient properties using semi-supervised variational autoencoder.

Abstract:

Determining the properties of chemical "in silico" (using computer) helps tasks such as drug discovery where large number of potential candidate molecules need to be screened for toxicity to human. Existing machine learning methods based on Variational Autoencoder (VAE) are able to predict properties and even generate novel molecules with certain salient properties (such as being safe to human).

However, determining the properties of chemicals in wet lab is a time-consuming and expensive endeavor, resulting relatively little labeled training data required for machine learning. Semi-supervised learning is a machine learning approach that leverages the data distribution of both labeled and unlabeled data.

In VAE, a hidden vector is a list of numbers that encodes all the information about a molecule; it can be used to reconstruct the molecule. Disentanglement of hidden vector ensures that the information about the property of a chemical is encoded at separate position compared to the structure of a chemical. With disentanglement in VAE, we can generate the chemical with a new property without changing its major structure, by manipulating only the "property position" in hidden vector.

In this project, we will modify the existing state of art approach for conditional generation of molecules (Junction Tree with Molecule Graph), to use the unlabeled portion of our data using semi supervised learning to improve property prediction performance. In addition, we'll disentangle the representation of chemical property in the hidden vector for conditional molecule generation task, if time permits.