Molecule Design and generation with Graph neural networks

Daoud Brahim
Nait Mohamed Aymen
Remili khalil
Djellab Ahmed

January 2, 2024

Summary of Research Papers on Graph-Based Models for Molecule Generation

Constrained Graph Variational Autoencoders for Molecule Design

Abstract and Introduction

- Introduces a variational autoencoder (VAE) model with graph-structured encoder and decoder, named Constrained Graph Variational Autoencoder (CGVAE).
- Incorporates gated graph neural networks (GGNNs) for molecule generation with domain-specific constraints.
- Addresses the challenge of sampling directly from a joint distribution over labeled nodes and edges.

Generative Model

- The generative process begins with latent vectors that form a specification for the graph to be generated. The process uses two functions: 'focus' to choose a node and 'expand' to decide on edges.
- 'Expand' is conditioned only on the partial graph structure and not the full generation history, to avoid deep computation graphs and training difficulty .

Training the Generative Model

- The model relies on a semantically meaningful latent space, with the encoder embedding each node into a latent space defined by mean μ_v and standard deviation σ_v .
- The decoder is supervised using generation traces extracted from graphs, focusing on valid expansions without penalizing for not reproducing sampled paths .

Optimizing Graph Properties

- The model facilitates the optimization of graphs with respect to numerical properties, using gradient ascent in the latent space.
- The overall training objective combines reconstruction loss, regularization on latent variables, and regression loss .

Application to Molecule Generation

- Specialized for generating chemical molecules, using datasets like QM9, ZINC, and CEPDB, with pre-processing to include specific bond types and node annotations ˙
- $Evaluatedonmoleculegenerationandoptimizationtasks, usingmetricssuchassyntacticvali$

- Extended to generate molecules with high Quantitative Estimate of Drug-Likeness (QED) values, using gradient ascent in latent space .

Conclusion

- CGVAE, with its sequential generative approach and enforcement of chemical rules, achieved state-of-the-art results in molecule generation and optimization.
- Highlights the need for collaboration with the chemistry community to develop additional metrics for real-world molecule design tasks .

A Two-Step Graph Convolutional Decoder for Molecule Generation

Abstract

- Proposes an auto-encoder framework for molecule generation, encoding molecular graphs into a continuous latent representation z, then decoding back to a molecule.
- Introduces a two-step decoding process: first generating a molecular formula, then placing bonds between atoms.
- Achieves a high reconstruction rate of 90.5% and reports the best property improvement results when optimization is constrained by molecular distance.

Introduction

- Discusses the challenges in drug discovery and material science of designing molecules with optimized chemical properties.
- Machine learning, especially deep learning, offers new avenues for learning molecular spaces for optimized molecule generation .

Molecule Auto-encoder

- Represents each molecule by a graph with vertices (atoms) and edges (bonds). The encoder generates a latent vector z from the molecular graph, which the decoder uses to recreate the molecule .

Molecule Encoder with Graph Neural Networks (GNNs)

- Uses GNNs to encode molecules of varying sizes N into fixed-size vectors d.
- Describes the encoding process with two steps: feature representation for all nodes and latent representation for the molecular graph .

Molecule Decoder

- Discusses two approaches for generating molecular graphs from a latent vector: auto-regressive models and one-shot models.
- The proposed model first generates all atoms and then all bonds in one shot, using a greedy beam search technique to ensure chemical validity.

Proposed Method

- The encoder reduces the molecular graph to a latent vector z. The decoder uses a multi-layer perceptron (MLP) for molecular formula generation and a graph convolutional network for bond classification .
- Atom generation is achieved by generating a molecular formula represented as a vector, which is then used to decide bond connections .
- Bond generation involves creating a fully connected graph and processing it with a graph convolutional network to predict bond types .
- Introduces positional features to differentiate atoms of the same type in the molecule .

Variational Auto-Encoder (VAE)

- Uses a VAE formulation to improve molecule generation, learning a parametrization of the molecular latent vector z.
- Total loss includes cross-entropy loss for edge and bag-of-atoms probability, and KullbackLeibler divergence for the VAE Gaussian distribution .

Beam Search

- Uses a greedy beam search technique for generating chemically valid molecules, addressing potential atom valency violations .

Experiments

- Conducts experiments on the ZINC molecule dataset.
- Reports significant improvements in reconstruction accuracy and the ability to generate new valid molecules.
- Includes property optimization, generating molecules with optimized chemical properties, and constrained property optimization, maintaining molecular similarity .

Conclusion

- Introduces an efficient VAE model for molecule generation, with a simpler implementation compared to previous techniques.
- Reports the highest reconstruction rate and best property improvement results for constrained optimization, demonstrating the model's ability to learn a good latent space representation of molecules .

Permutation-Invariant Variational Autoencoder for Graph-Level Representation Learning

Abstract

- Addresses the challenge of graph-level unsupervised representation learning, focusing on the complexity and permutation invariance of graphs.
- Proposes a permutation-invariant variational autoencoder for graph-structured data, which matches the node order of input and output graphs without imposing a particular node order or performing expensive graph matching.

Introduction

- Discusses the increasing interest in applying deep neural networks to non-Euclidean graph-structured data, with a focus on both supervised and unsupervised learning approaches.

Notations and Problem Definition

- Defines an undirected graph and its representation in matrix form, addressing the challenges posed by the permutation invariance of graphs.

Permutation-Invariant Variational Graph Autoencoder

- Describes the model's architecture, which includes an encoder, decoder, and a permuter model to align input and output graph node order.

- The permuter model is trained to predict the node order of the output graph, facilitating the alignment of input and output node orders.

Key Architectural Properties

- Highlights the permutation invariance of the graph embedding generated by the encoder, contrasting it with classical graph autoencoder frameworks.

Related Works

- Discusses existing research on unsupervised graph representation learning, focusing on node-level representation and deep methods based on Graph Neural Networks (GNNs).

Experimental Evaluation

- Details experiments on synthetic and molecular graphs from public datasets like QM9 and PubChem, demonstrating the model's effectiveness in graph reconstruction, class prediction, and property prediction.

Conclusion and Future Work

- Concludes with the proposal's significance as the first method for non-contrastive and non-adversarial learning of permutation invariant graph-level representations, highlighting its potential for graph generation and future scaling to larger graphs.