CSE60876 cover letter: 04/06/2023 (presentation date)
Kehan Guo

Paper 1: 3D Infomax improves GNNs for Molecular Property Prediction (*ICML 22*)

General Topic: "Molecular property prediction is one of the fastest growing applications of deep learning with critical real-world impacts. However, although the 3D molecular graph structure is necessary for models to achieve strong performance on many tasks, it is infeasible to obtain 3D structures at the scale required by many real-world applications."

Specific Behavior or Activity: "To tackle this issue, we propose to use existing 3D molecular datasets to pre-train a model to reason about the geometry of molecules given only their 2D molecular graphs. We show that 3D Infomax provides significant improvements for a wide range of properties, including a 22% average MAE reduction on QM9 quantum mechanical properties. Moreover, the learned representations can be effectively transferred between datasets in different molecular spaces."

Research Questions: The author is trying to encode this geometric information of molecule to the machine learning model in a computational effective manner.

Challenges: Using classical molecular dynamics simulations to explicitly compute a molecule's geometry before predicting its properties is computationally intractable for many real-world applications.

Paradigm: The authors present 3D Infomax, a graph neural network (GNN) pre-training solution that leverages 3D information to generate better learned embeddings and improve performance on down-stream prediction tasks where 3D information would be useful but not easily obtainable. The approach is useful for a range of downstream tasks involving molecules, including ones that are quantum mechanical, biological, and pharmacological in nature. They also demonstrate that the use of multiple 3D conformations (thereby encoding the inherent flexibility of molecules) further improves performance.

Problem: Recent methods to encode 3D geometric of a molecular graph requires giant computation cost which prevents GNN to encode these valuable features for accurate molecule property prediction.

Importance: Molecules are dynamic 3D structures that can exist in different spatial conformations. For a single 2D molecular graph, there are multiple low energy atom arrangements that are likely to occur in nature. These are called conformers and they can exhibit different chemical properties. For a model to properly capture 3D information, it is important to consider all the most likely conformations.

Claims: The author claims that by jointly using 2D and 3D structures of molecules to pre-train graph neural networks. During fine-tuning, the model can take 2D molecules as input. The pretraining phase ensures that the representations during fine-tuning contain latent 3D information.

State of Knowledge: As expected, most of the analyzed work uses ML techniques as their main method due to the inherent ability to adapt and learn autonomously.

Evidence: The author gathered empirical evidence from the quantitative analysis. The paper presents various graphs, pieces of code, and tables as evidence.

Story Structure: The author first explains why 3D information of molecule is important in molecule property prediction and identified that computing 3D features for every molecule is computationally intractable.
The authors then present 3D Infomax, a graph neural network (GNN) pre-training solution that leverages 3D information to generate better learned embeddings and improve performance on down-stream prediction tasks where 3D information would be useful but not easily obtainable. The approach is useful for a range of downstream tasks involving molecules, including ones that are quantum mechanical, biological, and

pharmacological in nature. They also demonstrate that the use of multiple 3D conformations (thereby encoding the inherent flexibility of molecules) further improves performance.

Paper 2: Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models

General Topic: "Since ChatGPT is trained with languages, it is currently not capable of processing or generating images from the visual world. At the same time, Visual Foundation Models, such as Visual Transformers or Stable Diffusion, although showing great visual understanding and generation capabilities, they are only experts on specific tasks with one-round fixed inputs and outputs.**"**

Specific Behavior or Activity: "To this end, we build a system called Visual ChatGPT, incorporating different Visual Foundation Models, to enable the user to interact with ChatGPT by 1) sending and receiving not only languages but also images 2) providing complex visual questions or visual editing instructions that require the collaboration of multiple AI models with multi-steps. 3) providing feedback and asking for corrected results.**"**

Research Questions: The author is trying to enable ChatGPT to process input of image modality.

Challenges: "Although powerful, ChatGPT is limited in its ability to process visual information since it is trained with a single language modality, while Visual Foundation Models (VFMs) have shown tremendous potential in computer vision, with their ability to understand and generate complex images"

Paradigm: The authors created Visual ChatGPT by integrating 22 Visual Foundation Models (VFMs) into ChatGPT, instead of developing a new multimodal ChatGPT from scratch. To facilitate this integration, they introduce a Prompt Manager that performs several functions, such as defining the input-output formats of each VFM, transforming visual data into language format, and managing the histories, priorities, and conflicts of various VFMs. By utilizing the Prompt Manager, ChatGPT can iteratively use these VFMs and receive their feedback until it meets the users' needs or fulfills the end condition.

Problem: "Could we build a ChatGPT-like system that also supports image understanding and generation?"

Importance: The injection of visual information to ChaGPT can help solve complex visual questions step-by-step.

Claims: The authors claims that with the help of the Prompt Manager, ChatGPT can leverage visual foundation models and receives their feedback in an iterative manner until it meets the requirements of users or reaches the ending condition

State of Knowledge: As expected, most of the analyzed work uses ML techniques as their main method due to the inherent ability to adapt and learn autonomously.

Evidence: The author gathered empirical evidence from the quantitative analysis. The paper presents various graphs, case studies and detailed analysis of each case generated by the model. In addition, the author discussed the limitation of ChatGPT in case of its model structure, capability and security.

Story Structure: The author begins the story by stating that although ChatGPT provides a language interface with distinctive conversational competency and reasoning capabilities across many domains, it is currently incapable of processing or generating images from the visual world. To overcome these limitations, the authors propose a system called Visual ChatGPT that incorporates different Visual Foundation Models to enable users to interact with ChatGPT using both language and images. The system can handle complex visual questions or instructions that require multiple AI models and steps, and also allows for feedback and corrections. To demonstrate the effectiveness of the proposed system, comprehensive experiments are conducted. In addition, the author did detail case study of the model and also discussed the limitation present.