MOLNET: A DEEP NETWORK FOR MODELING MOLECULES

A PREPRINT

October 24, 2022

ABSTRACT

Recently, transformer has proven to be an effective tool in chemical molecule analysis. In this paper, we propose MolNet: a model combines transformer and graph structure of the molecule. The key idea is that graph structure is used as an attention matrix which is multiplied to the node embeddings of the transformer. This method allows the attention score to perceive the structure of the molecular graph, and make more accurate predictions compared to the original transformer. The code of MolNet can be found in `https://github.com/zouxiaochuan/code_ogblsc2022`.

Keywords Graph Neural Network · Molecule Analysis · Transformer

1 Our Method

1.1 Feature Engineering

Molecular Features Atom and bond properties is obtained by the GetXXX(XXX is the property name of atoms or bonds such as FormalCharge, IsInRing, BondType, etc.) method of RDKit library.

Graph Structure Features Graph structure features play a crucial role in our model, every time we add a graph feature, we found the MAE is lower on the validation set. Graph structure features is represented as an attention matrix with the shape of $N \times N \times d$ where N is the number of atoms and d is the number of features. When calculating attention score, all features are mapped to an internal hidden dimension and then sum up. These structural features includes:

- Adjacent matrix: adjacent matrix is obtained directly from the link relationship of chemical bonds. When calculating the attention score, bond features obtained from 1.1 are mapped to an internal hidden dimension.
- Shortest path: the length of the path is used as a category feature.
- Count of same rings: this feature characterize how many rings between two atoms in the molecule, and is used as a category feature.

All the features are extracted by RDKit and NetworkX [Hagberg et al., 2008], it takes 6 minutes to finish on a 80-core machine for all 4m molecules of PCQM4M-V2 [Hu et al., 2021] dataset.

1.2 The Model

Our model is based on the idea of relative positions of transformer [Shaw et al., 2018].

In the following articles, if we do not specify otherwise, we will not distinguish between atom and node, bond and edge. Suppose $\mathbf{K}, \mathbf{Q} \in \mathbb{R}^{N \times h}$ is the key and query embeddings of the transformer and $\mathbf{A} \in \mathbb{R}^{N \times N \times h}$ is the graph feature extracted by 1.1, the attention score is carried out by the following equation:

$$attention_{i,j} = \mathbf{k_i} \cdot \mathbf{q_j^T} + \mathbf{k_i} \cdot \mathbf{a_{i,j}} + \mathbf{q_j} \cdot \mathbf{a_{i,j}} \tag{1}$$

where $\mathbf{k_i}, \mathbf{q_j}$ are i-th and j-th row of $\mathbf{K}, \mathbf{Q}$ respectively and $\mathbf{a_{i,j}}$ is the ij-th entry of $\mathbf{A}$. After the attention score is calculated, the rest of the model is the same as the standard transformer [Vaswani et al., 2017] and BERT [Devlin et al., 2019] model.

1.3 Training and Prediction

We tune our hyper parameters(e.g. epochs, batch size, etc.) on the validation set, then fix it and add validation set to training when we preparing our final model. We set the hidden dimensions of MolNet to 256, and number of layers to 24. It takes 50 hours to train on a single machine with 8 rtx3090 gpu cards.

In OGB-LSC at KDD Cup 2021 [Hu et al., 2021], ensemble methods proved to be very effective in improving test scores. So we ensemble 8 models each of which was trained on 98% of the whole training data, and then make average of the 8 models as our final score.

References

Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-2074. URL `https://aclanthology.org/N18-2074`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019.