MOLNET: A DEEP NETWORK FOR MODELING MOLECULES

A PREPRINT

Graph Learning Team
Department of CRO
Ant Group
HangZhou, China
xiaochuan.zxc@antfin.com

June 15, 2021

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. MolNet achieved 0.1244 on test set of PCQM4M dataset which is a top-3 winner of OGB-LSC at the KDD CUP 2021. The code of MolNet can be found in `https://github.com/zouxiaochuan/molnet_kddcup2021`.

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. Besides these properties, we found that atomic 3d coordinates is very useful in this task, we extract one 3d conformer and optimized by MMF method with the default setting.

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. Atoms and bonds in the path are mapped to an internal hidden dimension and sum them up.
- 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 two hours to finish on a 32-core machine for all 4m molecules. For test set, it takes about 4 hours to extract using a single core cpu.

1.2 The Model

We built our model based on the fantastic library transformers [Wolf et al., 2020], with a minor modification so that our model can incorporate graph structures. This is similar to the method used by BERT [Devlin et al., 2019] when dealing with relative positions.

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:

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

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 20. It takes 3 days to train on a single machine with 8 v100 gpu cards.

We found that the predicted values of the MolNet trained with different initial values are very different on the test set. The average absolute difference is about 0.04 with different initialization. Event with the same initialization and different running epochs, the difference is about 0.015. So we ensemble our model with 3 different versions of MolNet. Each version has slightly different hyper parameters, and we train 100 epochs and 150 epochs on each version. So we get 6 different predictions on the test set and then average them to get our final prediction.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

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.

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.