한국컴퓨터정보학회 하계학술대회 논문집 제30권 제2호 (2022. 7)

사전학습 전략과 딥러닝을 활용한 분자의 특성 예측

이승범O, 김지예*, 김동우*, 박재식*, 안성수*
O포항공과대학교 인공지능 대학원,
*포항공과대학교 인공지능 대학원
e-mail: {jk3472, dongwoo.kim, jaesik.park, sungsoo.ahn}@postech.ac.kr*, slee2020@postech.ac.krO

Molecular Property Prediction with Deep-learning and Pretraining Strategy

Seungbeom LeeO, Jiye Kim*, Dongwoo Kim*, Jaesik Park*, Sungsoo Ahn*
OGraduate School of Artificial Intelligence, POSTECH,
*Graduate School of Artificial Intelligence, POSTECH

● 요 약 ●

본 논문에서는 분자의 특성을 정확하게 예측하기 위해 효과적인 사전학습(pretraining) 전략과 트랜스포머(Transformer) 모델을 활용한 방법을 제시한다. 딥러닝을 활용한 분자의 성능을 예측하는 연구는 그동안 레이블이 부족한 분자데이터의 특성에 의해 학습 때 사용된 데이터이외의 분자데이터에 대해 일반화 능력이 떨어지는 어려움을 겪었다. 이 논문에서 제시한 모델은 사전학습(pretraining)을 수행할 때 자기지도학습(self-supervised training)을 사용하여 부족한 레이블에 의한 문제점을 피할 수 있다. 대규모 분자 데이터셋으로부터 학습된 이 모델은 4가지 다운스트림 데이터셋에 대해 모두 우수한 성능을 보여주어 일반화 성능이 뛰어나며 효과적인 분자표현을 얻을 수 있음을 보인다.

키워드: 사전학습(pretraining), 트랜스포머(Transformer), 분자표현학습(molecular representation learning)

I. Introduction

최근 머신러닝과 딥러닝과 같은 인공지능 기술은 특정 분야에 제한된 것이 아니라 생물학이나 화학과 같은 다른 여러 과학분야에서도 적극적으로 활용되고 있다. 신약개발을 위해 분자의 특성을 예측하는 분야도 대표적인 예 중 하나이다. 신약개발은 대표적인 고위험 고비용 산업분야 중 하나이다. 하나의 신약이 개발되기 위해서는 초기단계부터 승인을 얻기까지 걸리는 평균적으로 12년 이상의 시간이 걸리며 비용도 $1M 이상이 들기 때문이다 ([1], [2]). 최근 기계학습과 딥러닝과 같은 인공지능을 활용한 기술이 신약개발에 적극적으로 사용되고 있다. 약물의 후보군으로 사용되기 위한 물질의 특성을 값비싼 실험없이 사전에 예측 할 수 있기 때문이다. 하지만 특정 성질의 레이블이 적당한 크기를 갖는 데이터 셋으로 구축되기 위해서는 굉장히 많은 시간과 노력이 든다. 따라서 딥러닝 모델을 통해 학습을 효과적으로 하기 위해서 자기지도학습(self-supervised training)이나 기존에 있던 알고리즘을 사용하여 추출할 수 있는 값을 레이블로 사용하여 부족한 데이터 수 문제를 해결하고 이에 적합한 모델을 개발하는 연구가 활발히 진행 중이다.

본 논문에서는 분자의 구조를 잘 이해할 수 있는 학습 테스크와 트랜스포머 모델 구조를 활용해 높은 성능과 일반화 성능이 뛰어난 모델을 제시하였다 ([3]). 이 모델은 전통적인 기계학습 방법 (서포트 벡터 머신, 랜덤 포레스트)으로 얻어낸 분자 특성 예측 성능과 최신 딥러닝에서 나온 성능을 모두 뛰어 넘는 결과를 보여주었다.

II. Related Work

분자의 특성을 예측하는 모델을 개발하는 방법은 크게 사전학습전략과 모델을 개발하는 방식으로 많이 연구 되고 있다. 분자는 여러 원자와 이들 간의 결합관계로 이루어져 있기 때문에 원자(또는 node)와 분자(또는 그래프)로부터 효과적인 표현(representation)을 얻은 뒤 이 embedding 값을 예측하는데 사용하는 것이 주된 방법이다. 따라서 자기지도학습(self-supervised training)으로 학습을 할 때 node-level의 테스크와 graph-level의 테스크를 모두 사용하여 학습하는 것이 여러 테스크에 대해 성능이 더 좋다는 것이 알려져 있다

([4]). 이를 입증한 논문에서는 Graph Neural Networks (GNNs)에 효과적인 학습전략을 제시하면서 동시에 GNN의 모델 구조를 활용하여 분자의 특성을 예측한 모델들의 학습전략에 따른 성능을 제시하였다.

트랜스포머 모델은 NLP 분야에서 시작된 모델이지만 비전(vision)과 같은 다른 분야에서도 성공적인 성능을 보여주는 모델이다 ([3]). 따라서 이 모델을 활용하여 분자의 성능을 예측하려는 연구가 활발히 이루어지고 있다. Molecular Attention Transformer (MAT)는 분자의 성능을 예측하기 위해 테스크로 특정 원자(node)를 가린 뒤, 이 원자의 종류를 맞추는 테스크와 트랜스포머 모델 구조를 활용하였다 ([5]). 이때 분자의 구조에 대한 정보를 추가로 주기 위하여 기존의 트랜스포머의 self-attention layer에 원자간의 거리 정보와, 인접정보를 추가로 넣어주었다. GROVER라는 모델에서는 자기지도 학습 테스크로 분자의 부분구조를 가린 뒤, 이 구조를 맞추게 하는 방법을 새롭게 제시하였다 ([6]). 이때 학습에 사용된 모델 구조 역 트랜스포머이다. Graphormer 모델은 자기지도 학습 방법대신에 트랜스포머의 self-attention layer에 분자가 가질 수 있는 여러 추가적인 정보를 추가하였다. 예를들어 원자와 결합을 갖고 있는 다른 원자의 개수(degree), 한 원자에서 다른 원자까지 가기위한 최단거리(shortest path)등의 정보를 담을 수 있는 값을 추가로 넣어주어 학습을 진행하였다 ([7]).

III. The Proposed Method

이 섹션에서는 분자의 특성을 예측하기 위한 사전학습 전략과 캐스케이드 구조의 트랜스포머 모델에 대한 방법을 다룬다.

3.1 사전학습

제시 된 모델은 아래와 같이 총 5가지 종류의 사전학습 테스크(task)를 갖고 있다.

인접정보예측(Adjacency prediction): 인접정보 예측에서는 주어진 분자에서 두 원자사이의 연결정보를 예측하는 테스크이다. 연결정보의 유무만을 예측하는 것으로 손실함수로는 binary cross entropy를 사용하였다.

본드타입예측(Bond-type prediction): 본드타입 예측이란 두 원자 사이의 연결이 어떤 종류인지(예시: 단일결합, 아로마틱 결합 등) 예측하는 테스크이다. 총 5가지의 특성(feature)를 갖고 있으며 cross entropy가 손실함수로 사용되었다.

원자예측(Atom-prediction): 원자예측이란 주어진 원자가 어떤 원자인지를 예측하는 테스크이다 이 모델에서는 하나의 분자를 이루고 있는 모든 원자의 종류를 예측해야 한다. cross entropy가 손실함수로 사용되었다.

Functional-group 예측: functional-group 예측이란 한 분자내에 들어있는 특정 부분구조 (예, ring 구조)를 예측하는 테스크이다. 이때 부분구조는 사전에 정의된 167가지의 functional group에 해당되는 구조이다. RDkit [(10)]을 통해 추출되었으며 cross entropy가 손실함수로 사용되었다.

분자특성예측 (property prediction): 분자가 갖고있는 성질(예: water solubility)을 예측하는 구조이다. 200가지의 연속적인 값을 갖는 수치로 Rdkit을 통해 추출되었다. Mean Squared Error(MSE)가 손실 함수로 사용되었다.

3.2 모델구조

제시된 모델은 캐스케이드 구조를 갖고 있다. 기존의 연구들과 다르게 이 트랜스포머 구조는 두 종류의 인코더인 구조 인코더(structure encoder)와 특성 인코더(property encoder)를 갖고 있으며, 각 인코더마다 계층적으로 서로 다른 손실함수가 배정되었다. 구조 인코더에서는 분자의 구조정보를 학습하고, 이때 학습된 구조정보를 바탕으로 특성 인코더에서 분자의 특성과 관련된 예측을 수행하는 것이 이 모델의 핵심이다.

구조 인코더(structure encoder): 3.1에서 제시한 테스크 중 분자의 구조에 해당되는 인접정보예측, 본드타입예측, 원자예측이 이 구조 인코더를 통과한 후 수행된다. 이때 구조 인코더를 통과한 후 node embedding이 각각의 테스크마다 독립적인 다중 퍼셉트론 (multi-layer perceptron, MLP)을 통과한후 테스크를 수행하였다. 따라서 이 손실함수를 통해 나오는 gradient들은 구조 인코더 이후에 있는 layer에는 전달되지 않고 이전에 있는 layer에 만 전달되어 weight가 업데이트 된다.

특성 인코더(property encoder): 특성 인코더에서는 나머지 테스크인 functional group 예측, 분자특성 예측이 수행된다. 이 테스크를 사용하기 위해 구조 인코더 이후 하나의 토큰(token)을 추가되었고, 이 토큰이 특성 인코더를 통과한 후 얻어진 embedding값이 손실함수를 예측하는데 사용되었다. 이 토큰은 분자 전체단의 특성을 예측하기 위해 사용 되었다.

이외에 Laplacian Positional Encoding(LPE)가 추가로 input 단계에서 node의 embedding값에 추가 되었다 ([9]). 이는 분자의 구조정보를 간접적으로 넣어주기 위한 inductive bias로서, 기존 트랜스포머 모델의 positional encoding에 사용되는 기법 중 하나이다.

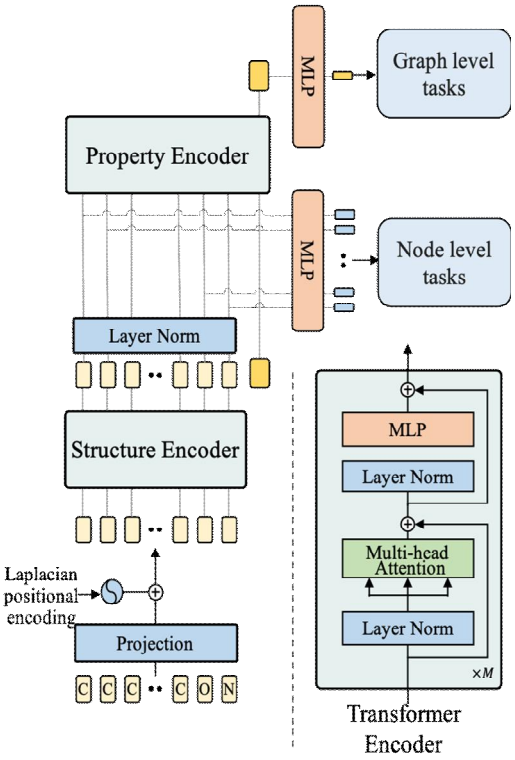

Fig. 1. Model architecture

IV. Experiment

이 모델을 학습시키기 위해 ChEMBL과 Pubchem에서 약 2백만 개의 분자데이터 셋을 추출하였다 ([11], [12]). 이 대규모 데이터셋을 통해 사전학습된 모델의 성능을 다른 4가지 다운스트림 테스크 (downstream task)에 미세조정 한 후 결과값을 측정하였다.

Table 1. Downstream dataset

데이터셋	크기	Task type
BBBP	2,039	Classification
ESOL	1,128	Regression
FreeSolv	642	Regression
Estrogen	1,961	Classification

각 데이터셋은 특정 성질에 대한 레이블을 갖고 있다. (예: BBBP 데이터셋은 분자가 혈액뇌장벽 (blood-brain barrier)을 통과할 수 있는지에 대한 여부)

Table 2. Result

모델	ESOL (↓)	Freesolv (↓)	BBBP (↑)	Estrogen (↑)
SVM	.479(0.5)	.461(.07)	.723(.01)	.772(.01)
RF	.534(.07)	.524(.09)	.721(.01)	.791(.01)
GCN[8]	.369(.32)	.299(.68)	.695(.13)	.730(.03)
MAT[5]	.278(.02)	.265(.04)	.737(.09)	.773(.01)
GROVR [6]	.303(.04)	.270(.03)	.726(.07)	.758(.06)
Graphormer[7]	.273(.01)	.271(.05)	.725(.01)	.805(.34)
Ours	**.252(.01)**	**.242(.04)**	**.754(.01)**	**.820(.01)**

이 실험결과에서 데이터셋에 있는 (↑),(↓) 표시는 테스크 종류에 따른 성능 기준을 말한다. 예를들면 (↓)는 낮을수록 우수하다는 것을 의미한다. SVM은 support vector machine, RF는 random forest를 통해 학습 된 모델이다.

다음은 실험 결과이다. Table 2 결과값에서 regression task의 경우 Mean Absolute Error(MAE)가 성능 측정의 지표로 사용되었으며, classification task의 경우 ROC-AUC CURVE가 평가지표로 사용 되었다. Table 2 에서 볼 수 있듯이 현재 제시된 방법의 사전학습 테스크를 사용한 트랜스포머 구조의 모델이 굉장히 좋은 성능을 보였음을 확인 할 수 있다. 이 트랜스포머 모델은 self-attention layer에 특정한 변화를 주지 않고 인코더만 두 부분으로 나눈 뒤 특정 태스크를 분리하는 전략을 사용하였다. 이는 우리 모델구조가 효과적인 사전학습 전략을 사용하였으며 주어진 손실함수를 적절히 잘 배치하였음을 보여준다.

V. Conclusion

본 논문을 통해 인공지능을 활용하여 분자의 특성을 예측하는 모델을 제시하였다. 효과적인 사전학습 테스크들과 이 테스크를 적절히 활용하는 트랜스포머 모델의 제안을 통해서 인공지능의 기술이 화학이나 생물과 같은 다른 분야에서도 적극적으로 활용 될 수 있는 점을 확인 할 수 있었다.

본 논문에서 제시한 방법은 (1) 대규모 데이터셋을 통해 사전학습을 진행하며, (2) 학습에 자기지도학습 통해 레이블일 부족한 문제를 극복하고, (3) 손실함수를 적절히 배치하여 두가지 종류의 인코더를 가지는 트랜스포머 모델구조를 구축 하는 것이다. 학습된 모델은 다른 downstream dataset에 대해 좋은 성능을 보였기 때문에 일반화 능력이 우수함을 알 수 있다. 따라서 이 모델에서 학습된 방법을 통해 논문에서 제시된 데이터 이외에도 다른 분자데이터셋에서도 좋은 성능을 보일 것을 기대 할 수 있다.

REFERENCES

[1] Mohs, R. C. and Greig, N. H. Drug discovery and development: Role of basic biological research. Alzheimer's &Dementia: Translational Research & Clinical Interventions, 3(4):651–657, 2017.

[2] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. Principles of early drug discovery. British journal of pharmacology, 162(6):1239–1249, 2011.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[4] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for Pre-training Graph Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.

[5] Maziarka, Ł., Danel, T., Mucha, S., Rataj, K., Tabor, J., and Jastrz˛ebski, S. Molecule attention transformer. arXiv preprint arXiv:2002.08264, 2020

[6] Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. GROVER: Self-supervised Message Passing Transformer on Large-scale Molecular Data. In Advances in Neural Information Processing Systems (NeurIPS), 2020

[7] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do Transformers Really Perform Bad for Graph Representation? In Advances in Neural Information Processing Systems (NeurIPS), 2021.

[8] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J.,Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A.,and Adams, R. P.Convolutional networks on graphs for learning molecular fingerprints. Advances in Neural Information Processing Systems (NeurIPS), 2015.

[9] Veli˘ckovi´c, P., Cucurull, G., Casanova, A., Romero, A.,Lio, P., and Bengio, Y. Graph attention networks. Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[10] Landrum, G.Rdkit: Open-source cheminformatics software. 2016. URL https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

[11] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, YvonneLight, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research, 40(D1),09 2011.

[12] Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. PubChem substance and compound databases. Nucleic acids research, 2016.