COSE474-2024F Final Project Proposal

장정현

December 8, 2024

1 Introduction

Motivation

LUMO-HOMO gap은 전자가 채워진 가장 높은 분자 오비탈과 전자가 없는 가장 낮은 오비탈 간의 에너지 준위 차이로, 반도체 소자뿐만 아니라 유기 소재의 반응성, molecular simulation에서의 DFT 계산 등 화학 분야에서 굉장히 유용한 molecular property이다. Gap을 예측하는 것은 소재 및 공정 개발의 시간 및 비용을 절감한다.

ChemBERTa는 SMILES로 표현한 분자의 화학적인 property 예측에 적합한 pre-trained model이며, SMILES는 분자 구조를 프로그램이 해석하기 쉽도록 텍스트로 표기한 것으로, 동일한 구조에 대해 여러 SMILES 표기가 존재할 수 있다. Canonical SMILES는 하나의 구조에 대해 하나의 표현만이 가능하도록 규칙화된 SMILES이다. 이는 AI가 분자 구조에 대해 더 일관된 패턴을 학습할 수 있도록 하지만, 그 영향이 명확한 것은 아니다. 따라서 실제 실험을 통해 Canonical SMILES가 training data로서 SMILES에 비해 갖는 이점을 확인할 필요가 있다.

Problem Definition

본 연구에서는 ChemBERTa 모델을 LUMO-HOMO gap 예측을 위해 fine-tuning하고, SMILES와 canonical SMILES를 입력으로 사용할 때 예측 성능을 비교하여 분자 표현 방식의 차이가 학습 및 일반화 성능에 미치는 영향을 분석한다.

Concise Description of Contribution

ChemBERTa 모델을 LUMO-HOMO gap 예측을 위해 fine-tuning 한다. 또한 그 과정에서 canonical SMILES와 같은 분자 표현 전처리의 영향을 실증적으로 제시한다.

2 Methods

Significance/Novelty

LUMO-HOMO gap 은 재료과학, 물리화학 등 다양한 분야에서 활용되는 property이며, 이를 예측하는 fine-tuning된 모델은 반도체, 약품 개발 등에 유용하게 활용될 수 있다. 또한 분자 구조 표현에서 SMILES가 갖는 다양성을 제거하는 것이 (canonical SMILES) 모델의 성능에 미치는 영향을 정량적으로 분석하는 것은, 화학자들이 ChemBERTa와 같은 모델을 사용할 때 입력의 전처리에 대한 유용한 가이드라인이 될 수 있다.

ChemBERTa 모델 학습 과정은 아래와 같이 설계되었다.

1. 데이터 준비
   - QM9 데이터셋 준비
   - Tokenizer를 통해 각 SMILES의 token_length 확인하여 적절한 max token length 계산
   - RDKit을 통해 QM9 데이터셋을 Canonical SMILES로 변환 및 기존 SMILES가 Canonical이 아닌 개수 계산
   - Training 및 test dataset 분리
2. Fine-tuning
   - ChemBERTa 모델을 기반으로 출력 레이어를 회귀 문제에 맞게 변경
   - SMILES 및 Canonical SMILES dataset에 대해 training 수행
3. 성능 확인
   - MSE(Mean Squared Error)를 사용하여 두 입력 형식(SMILES, Canonical SMILES)의 예측 성능을 비교

3 Experiments

3.1 Dataset

본 연구에서 사용한 QM9 데이터셋은 133,886개의 유기 분자에 대해 LUMO-HOMO gap을 포함한 화학적 property 정보를 제공한다.

- **(Non-canonical) SMILES 데이터**: QM9 데이터셋에 포함된 원본 SMILES.
- **Canonical SMILES 데이터**: Non-canonical SMILES 데이터를 RDKit 라이브러리를 활용하여 Canonical SMILES 형식으로 변환.

Train/Test Split:

- 데이터셋은 90:10 비율로 분리.
- Train 데이터: Non-canonical 및 Canonical SMILES 각각 약 120,000개.
- Test 데이터: Non-canonical 및 Canonical SMILES 각각 약 13,000개.

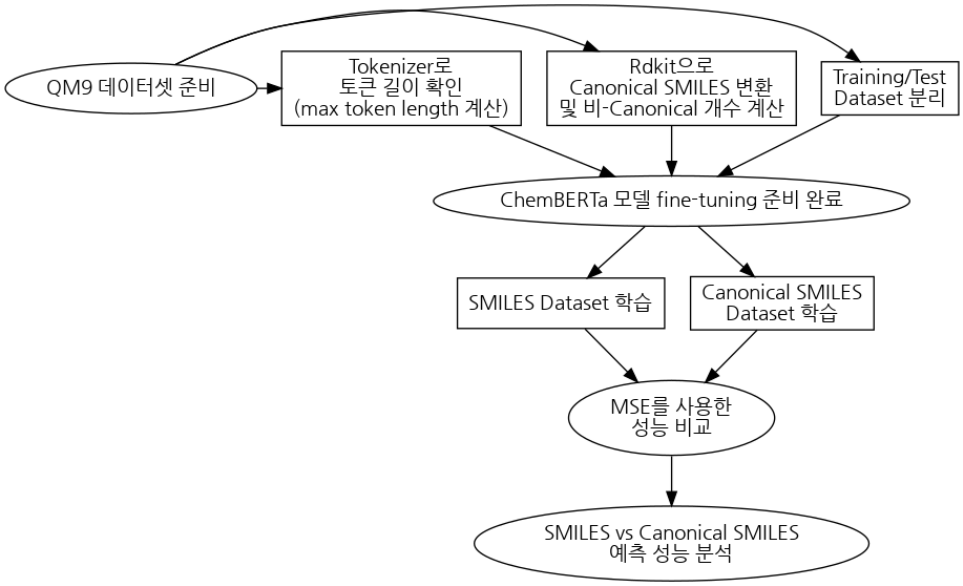

Figure 1: Work flow figure

3.2 Computing Resources

환경:

- Google Colab (무료)

Hardware:

- GPU: NVIDIA Tesla T4
- CUDA Version: 12.2
- CPU: Intel Xeon 2.00GHz

Software:

- Operating System: Ubuntu 22.04.3 LTS
- Python Version: 3.10.12
- PyTorch Version: 2.5.1+cu121
- Other Packages:
  - transformers==4.46.3
  - numpy==1.26.4
  - pandas==2.2.2
  - rdkit==2024.3.6

3.3 Experimental Design

실험 목표: Canonical SMILES와 Non-canonical SMILES를 입력으로 사용할 때, ChemBERTa 모델의 LUMO-HOMO gap 예측 성능 차이를 비교한다.

모델 설정:

- **ChemBERTa**: seyonec/ChemBERTa-zinc-base-v1 사용.
- 회귀 문제에 맞게 출력 레이어 수정.
- Learning rate: 5e-5
- Batch size: 16
- Epoch: 2
- Loss function: Mean Squared Error (MSE)

3.4 Quantitative Results

SMILES Type	MSE
Non-Canonical SMILES	0.0003443
Canonical SMILES	0.0002960

Table 1: Canonical SMILES와 일반 SMILES의 MSE 비교

3.5 Qualitative Results

각 SMILES를 기반으로 한 HOMO-LUMO gap 예측에 대한 fine-tuning이 모두 성공적으로 수행되었다.

3.6 Analysis

Canonical SMILES가 test 결과 약 14.03% 낮은 MSE를 보여 HOMO-LUMO gap 예측 성능이 상대적으로 우수하다. 따라서 분자의 chemical property를 예측하기 위해 ChemBERTa 기반 모델을 사용할 때, SMILES input을 Canonical하게 전처리한 후 활용하는 것이 유용한 것으로 확인된다.

3.7 Why the Proposed Method is Successful

Canonical SMILES 사용 시 성능이 더 우수한 이유는 다음과 같이 추론할 수 있다.

- Canonical SMILES는 분자의 고유한 표현을 제공하므로, 데이터 중복 문제를 줄여 모델 학습을 효율적으로 만든다.
- 일반 SMILES는 동일한 분자를 다양한 방식으로 표현할 수 있어 모델이 불필요한 복잡성을 학습할 가능성이 크다.

그러나, 일반 SMILES의 성능이 낮아도 다양한 표현 학습으로 일반화 성능을 높이는 방향으로 활용할 여지가 있을 수 있다.

3.8 Future Direction

- **다른 데이터셋에 대한 실험**: Canonical SMILES의 성능 이점이 다른 대규모 데이터에서도 유지되는지 확인.
- **다양한 모델 평가**: ChemBERTa 외에 다른 사전학습 모델을 활용.
- **다양한 화학적 특성**: LUMO-HOMO gap 외에 다른 물리화학적 특성 예측 확장.

References

[1] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

[2] Ingoo Lee and Hojung Nam. Infusing linguistic knowledge of smiles into chemical language models, 2022.