GPT 모델을 이용한 분자 구조 생성

김도연

CONTENTS

기존 분자구조 생성 연구

사전 지식

생성 모델과 성능 평가 지표

생성 모델 : 주어진 training data와 같은 distribution을 가진 새로운 sample을 만들어내는 모델

- sampling : training data에 존재하지는 않으나 가장 비슷한 분포를 나타내는 임의의 sample 추출

- **Validity** : (number of valid molecules) / (number of generated samples)
- **Uniqueness** : (number of unrepeated molecules) / (number of valid molecules)
- **Novelty** : (number of molecules not included in the training set) / (number of unique molecules)
- **Diversity** : 생성된 분자들의 diversity(mode collapse나 similarity 측정)
- **FID**(Frechet Inception Distance)

$$FID = |\mu_T - \mu_G|^2 + Tr(\Sigma_T + \Sigma_G - 2(\Sigma_T\Sigma_G)^{1/2})$$

* T = test set, G = generated image set

기존 분자구조 생성 연구

분자 구조 생성의 주요 아이디어

분자들의 분포를 추정하여 target properties에 대한 새로운 분자를 생성 및 sampling

VAE Latent variables 분포의 불충분한 예측으로 **validity가 낮음.**

GAN 이산화된 variable로 표현된 데이터에 대해 **diversity가 낮음.**

기존 분자구조 생성 연구

“ Molecular Generative Model Based on an **Adversarially Regularized Autoencoder** ”

- **Encoder-Decoder 구조**
- **Latent variable model**

이산화된 분자 구조를 input으로 넣으면 연속적인 latent representation으로 변형

- **Key distinct feature** from **GAN**

→ GAN의 adversarial training을 Latent variable의 분포를 추정하는데 사용

- **성능 지표** : validity, uniqueness, novelty, diversity(Tanimoto similarity)
- 특정 properties에 따라 분자를 생성하는 **CARAE**(Conditional ARAE)로도 발전

기존 분자구조 생성 연구

분자 생성과 관련된 Benchmark dataset

MOSES
(Molecular Sets)

분자 생성 모델링
생성, 학습

GuacaMol

조건에 맞는 분자 디자인 및 최적화
성능 평가

기존 분자구조 생성 연구

MOSES dataset

Baseline으로 사용한 모델

- Character-level Recurrent Neural Network (CharRNN)
- Variational Autoencoder(VAE)
- Adversarial Autoencoder(AAE)
- Junction Tree Variational Autoencoder (JTN-VAE)
- Latent Generative Adversarial Network(LatentGAN)

최근에 MOSES dataset을 활용한 모델/논문

- **Generative Molecular Transformer(GMTransformer)** (2023.09.)
- **Generative Pre-trained Transformer(GPT) based model with relative attention for de novo drug design** (2023.10.)

기존 분자구조 생성 연구

MOSES dataset

- **Internal Diversity(IntDiv_p)**
  - **목적** : chemical diversity 측정 → 모델이 사용하는 chemical space 범위 나타냄.
  - Tanimoto similarity 사용하여 SMILES 간 유사성 표현 by RDkit

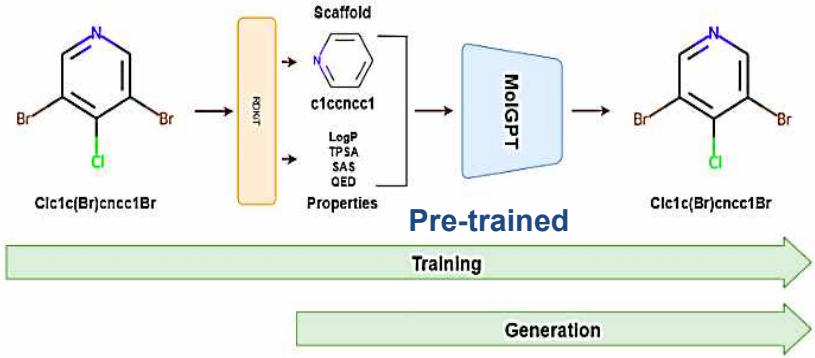
$$\text{IntDiv}_p(S) = 1 - \sqrt[p]{\frac{1}{|S|^2} \sum_{s1,s2 \in S} T(s1, s2)^p}$$

  → mode collapse와 같은 실패 케이스 탐지

- **Frechet ChemNet Distance(FCD)**
  - **목적 :** 모델이 얼마나 통계적 정보를 잘 capture하는지 측정(chemical & biological properties)
  - FID의 Inception 모델 대신 ChemNet 모델 사용해 feature 간의 거리 측정

$$\text{FCD}(G, D) = \|\mu_G - \mu_D\|^2 + Tr(\Sigma_G + \Sigma_D - 2(\Sigma_G\Sigma_D)^{1/2})$$

MolGPT와 Fine-tuning

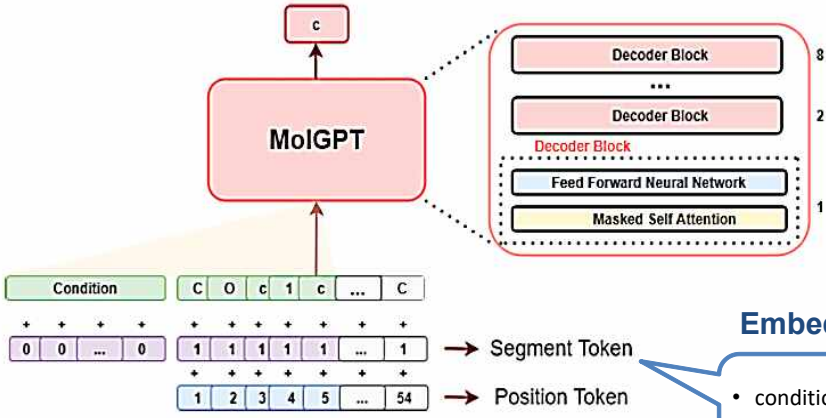


분자 생성과정

- **condition data의 start token을 모델에 제공**
  - SMILES training set에서 첫번째로 발생하는 token들의 list에서 **weighted random sampling**
  - **weight** : SMILES training set의 첫번째 위치에서 발생하는 빈도에 의해 결정

  SMILES tokenizer 반복적 사용
- **start token부터 시작해 순차적으로 다음 token 예측하여 분자 생성**
- **property condition도 같이 제공하여 분자 sampling**

MolGPT와 Fine-tuning



- **Transformer-based**

  + Auto-regressive decoder
- target-specific embedding을

  Multi-head attention의 value & key로 사용

- 8 decoder blocks + multiple masked SA 적용

Embedding

- conditional training에 사용
- 특정 input이 condition의 token인지 molecules SMILES token인지 구별

MolGPT와 Fine-tuning

Fine-tuning

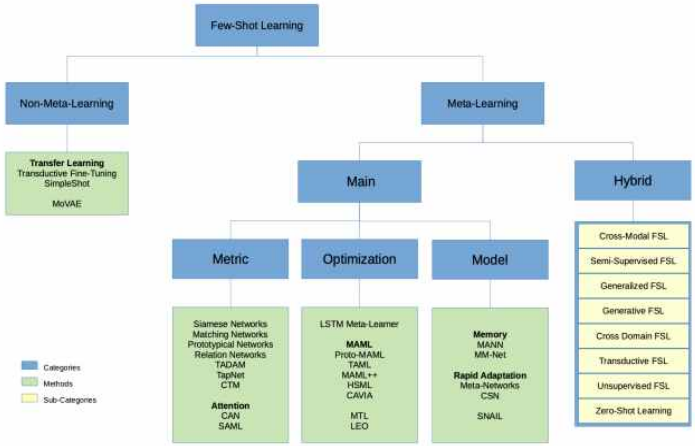
- 기존에 사전 학습된 모델을 기반으로 새로운 task에 맞게 변형하고 이미 학습된 모델의 가중치를 미세하게 조정하여 학습시키는 방법
- property-condition (logP, SAS, TPSA, QED)을 RDkit으로 측정, 추출 후 fine-tuning에 사용

Few-shot learning

소량의 데이터(few-shot)으로 trained data와의 유사성을 학습시키는 방법

MolGPT와 Fine-tuning

Few-shot learning ▶ Non-Meta learning ▶ Transfer learning

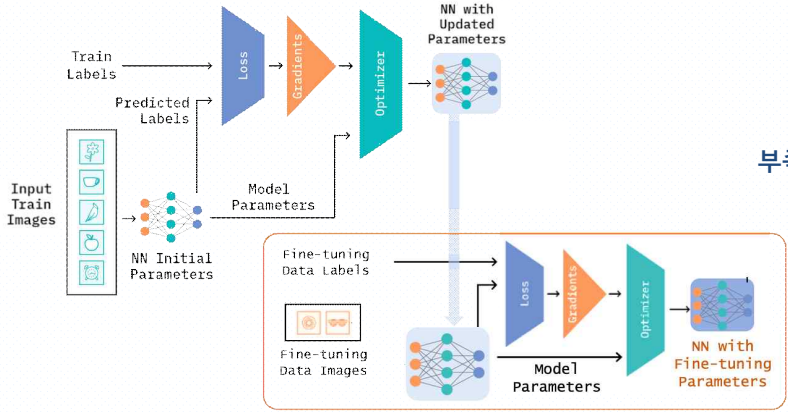


- **pre-training : 대량의 데이터로 사전 학습 모델 생성**
- **사전 학습 모델 weight를 그대로 가져온 뒤 각 Task에 맞는 데이터(Fine-tuning 데이터) 로 재학습해서 모델 생성**
- **inference 과정에서 conditioning으로 이용할 수 있는 약간의 task에 대한 설명이 주어지지만, 직접 학습에 활용하지는 않는다.**
- **task에 대한 설명과 함께 task에 대한 K개의 example들 제공**

(K : model의 context window, 대략 10~100의 값.)

MolGPT와 Fine-tuning

효과적인 Fine-tuning 방법



- **dense layer 부분에서 일부 layer들의 weight parameter freeze**

부족한 sample 수로 인한 overfitting의 가능성을 줄여야 한다!

- **오버피팅을 줄이기 위해...**
  - **data augmentation**
  - **정규화**
  - **적당한 크기의 epochs**

새로운 분자 구조 생성 모델 구현 사항

필요한 dataset

Data-imbalance 문제

사용할 성능 평가 지표

데이터 전처리(SMILES 표현)

Pre-training & Fine-tuning

.....

새로운 분자 구조 생성 모델 구현 사항

필요한 dataset

- **Tox21 (12,060 training samples & 647 test samples)**
  - 화합물의 독성에 대한 데이터 → 독성 스크리닝
  - 핵 수용체(Nuclear Receptor)와 스트레스 반응 경로와 관련된 독성에 초점
- **MOSES (Molecular Sets) (1.6M training samples & 176k test samples)**
- **GuacaMol**

새로운 분자 구조 생성 모델 구현 사항

Data imbalanced condition → **Resampling** 기법으로 해결 가능!

- **Data-imbalanced** : Tox21 dataset에서 toxic 화합물의 수 <<< nontoxic 화합물의 수 - 이진 분류

→ 낮은 accurcy + 0.5보다도 낮은 sensitivity

→ toxicity hazard에 대한 평가가 어려워진다. (minority class 인식이 어려워 분류 경계가 biased됨)

- **Minority data를 oversampling하는 resampling 기법**
  - random shifting + augmentation

참고 : The Effect of Resampling on Data-imbalanced Conditions for Prediction towards Nuclear Receptor Profiling Using Deep Learning(2020)

새로운 분자 구조 생성 모델 구현 사항

사용할 성능 평가 지표

- **Validity** : (number of valid molecules) / (number of generated samples)
- **Uniqueness** : (number of unrepeated molecules) / (number of valid molecules)
- **Novelty** : (number of molecules not included in the training set) / (number of unique molecules)
- **Internal Diversity** : 생성된 분자들의 diversity(mode collapse나 similarity 측정)
- **Frechet ChemNet Distance(FCD)** : ChemNet 모델의 penultimate layer에서 얻은 molecular features로 측정
- **KL divergence**
- **fragment similarity**

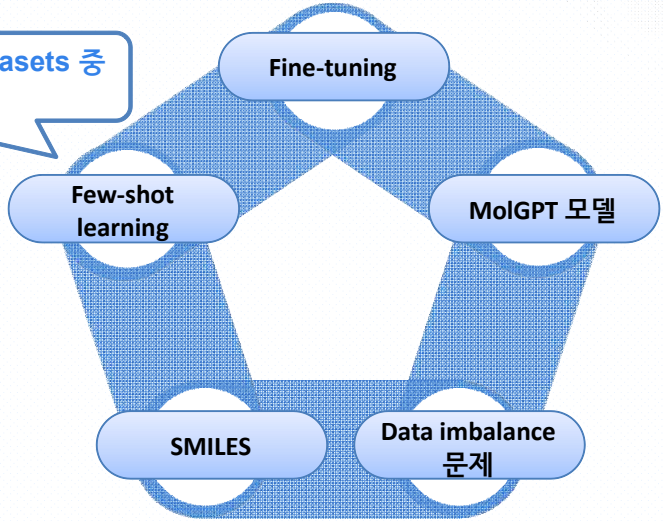
새로운 분자 구조 생성 모델 구현 사항

데이터 전처리

- **SMILES 토큰화**
  - SMILES training set에서 첫번째로 발생하는 token들의 list에서 weighted random sampling
    - weight : SMILES training set의 첫번째 위치에서 발생하는 빈도에 의해 결정
  - **Embedding** : Position value embedding + **Segment** token embedding (conditional)
- **SMILES vocabulary 생성**
  - 100%에 근접한 validity를 위해 SMILES 형식에 맞는 vocabulary를 만들어놓고 sampling

앞으로의 연구 방향성

MoleculeNet의 benchmark datasets 중 Tox21 dataset 이용



앞으로의 연구 방향성

새로운 Fine-tuning 방식인 PEFT도 알아보자!

```
RuntimeError: CUDA out of memory. Tried to allocate 200.00 MiB (GPU 0; 15.78 GiB total capacity; 14.56 GiB already allocated; 38.44 MiB free; 14.80 GiB reserved in total by PyTorch) If reserved memory is >> allocated memory try setting max_split_size_mb to avoid fragmentation. See documentation for Memory Management and PYTORCH_CUDA_ALLOC_CONF
```

대규모의 모델을 학습시킬 때는 많은 양의 GPU 자원 필요!

앞으로의 연구 방향성

새로운 Fine-tuning 방식인 PEFT도 알아보자!

PEFT란?

Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware

1. **LoRA : Low-RANK Adaptation of Large Language Models**
2. **Prefix Tuning**
3. **Prompt Tuning**
4. **P-Tuning**

앞으로의 연구 방향성

새로운 Fine-tuning 방식인 PEFT도 알아보자!

PEFT란?

Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware

1. LoRA : Low-RANK Adaptation of Large Language Models

pre-trained model의 weight의 일부를 freeze하는 것 대신에
훈련 가능한 rank decomposition 행렬 (low-rank 행렬) 추가
→ adaptation 동안 변화하는 dense layer의 rank 행렬을 최적화

Q & A

References

- **Molecular Generative Model Based on an Adversarially Regularized Autoencoder(2020)**
https://pubmed.ncbi.nlm.nih.gov/31820983/

- **MolGPT: Molecular Generation Using a Transformer-Decoder Model(2021)**
https://pubs.acs.org/doi/10.1021/acs.jcim.1c00600

- **The Effect of Resampling on Data-imbalanced Conditions for Prediction towards Nuclear Receptor Profiling Using Deep Learning(2020)**
https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201900131?casa_token=Z2UXtKX5KxoAAAAA%3A0DGFx0olrHAcCtuvKu8ccP0NnBDOPmU2VmqAcT-Cm_Di8l_a6lMgrZGoTX04b1U7uJZnSlaEabNHdyHyXA

- **데이터의 신비한 변신: 웹 스크래핑과 ChatGPT Fine-tuning의 예술 - 파인튜닝 과정**
https://wikidocs.net/198965

- **PEFT: Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware(2023)**
https://huggingface.co/blog/peft?fbclid=IwAR2hc-x_oaQ5e4vitAJMoC-BaGxyKyPJ-oKgU2pugQsCpNF3mq7mT3UUydE&utm_source=pytorchkr

Thank You