목차

1 과제 배경 및 목표

1.1 과제 배경

시대가 변하면서 사람들의 평균 연령 뿐만 아니라 건강에 대한 의식 또한 점점 높아지고 있다. 예전에는 치료가 안되던 불치병들도 속속들이 치료법이 나타나고 있으며 사람들 또한, 점점 웰빙을 추구하는 경향이 나타나고 있다. 하지만 아직까지도 에이즈, 파킨슨병, 말기 암 등의 사망률이 높거나 심각한 질병들의 경우 불치병으로 자리잡고 있고 의학기술이 늘어남에 따라 희귀 질환 또한 늘어나고 있다. 따라서 이러한 질병의 추세에 발 맞춰 신약개발에 대한 연구에 대한 필요성이 점차 늘어나고 있다.

1.2 과제 목표

본 졸업 과제는 딥러닝 기반 모델을 통해 drug와 target의 sequence 정보만을 이용한 효과적인 Drug-Target interaction binding affinity 예측을 목표로 합니다.

- 데이터셋
- 데이터 전처리
- 모델 구현
- 서비스 구현

.

2 요구 조건 분석

2.1 데이터셋

- Binding Affinity 예측 평가로 사전에 사용된 Davis Kinase dataset와 KIBA dataset등을 이용하여 만든 모델의 평가로 쓴다.

2.2 데이터 전처리

- Integer/label 인코딩 기법을 이용하여 단백질의 원소(탄소(C), 수소(H), 산소(O)등)에 대한 라벨링을 숫자 1,2,3으로 표현한다. 또한, 단백질 서열의 다양한 길의 표현을 고려하여 최대 표현 길이 제한 등을 둔다.

2.3 모델 구현

- Protein - ligand 상호작용에 대한 예측을 regression 문제로 접근하여 binding affinity를 평가한다.
- Convolutional Neural Network(CNN)을 도입하여 필터 사이의 local dependency를 적극 활용한다.

2.4 서비스 제공

- 결과를 보기 쉽게 정리하기 위하여 웹서비스 혹은 로컬 서비스에 만들어진 모델에 대한 결과물 및 평가를 제시한다.

3 현실적 제약 사항 및 대책

3.1 제약 사항

- 사용할 수 있는 Dataset 종류가 한정적이고 병원과의 협조가 불가피한 상황이다
- Dataset bias와 generalization 문제가 있다
- Data Labeling 과정이 시간이 많이 필요한 작업이다
- CNN은 black-box 모델이기 때문에 binding affinity 예측을 결정하는 원인들을 해석하기 어려울 수 있다
- 데이터 해석을 위한 컴퓨터 공학적인 측면의 발전에 아직 미세 분자 구조에 대한 정보 해석이 따라오지 못하여 좋은 모델의 예측에도 불구하고 부정확한 결과가 탄생할 수 있다.

3.2 대책

- 송길태 교수님의 AI랩의 Data labeling이 완료된 데이터를 이용한다
- Data augmentation을 통해 제한된 dataset의 크기를 늘린다
- Drug-Target 반응의 더 깊은 이해를 위해 다중 오믹스 data를 이용한다
- Stratified sampling, cross-validation을 통해 모델 성능에 대한 더 정확한 평가를 얻을 수 있다.

4 설계 문서

4.1 개발 환경

- 개발 언어
  - Python(데이터 전처리, 모델 구현), Keras(신경망 구현)
- 개발 도구
  - TensorFlow(기계 학습)
- 실행 환경
  - GPU (with cuDnn)

4.2 사용 기술

- Kronecker Regularized Least Squares (KronRLS)

$$J(f) = \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda ||f||_k^2$$

  - KronRLS을 사용하는 목적은 data-fitting term과 regularization term의 합에 대한 최솟값을 구하는 데에 있다. Data-fitting 파트는 prediction function을 통해 얻은 예측값들과 참값에 대한 오차를 측정하기에 이를 줄여야 prediction function이 training data set을 잘 반영한다고 볼 수 있다. 또한, regularization 파트는 prediction function의 복잡도 혹은 smoothness에 대한 penalty를 측정하기에 overfitting을 방지하고 미학습 data에 대한 모델의 성능을 증가하기 위해 이 또한 줄여야 할 요소이다.

- SimBoost

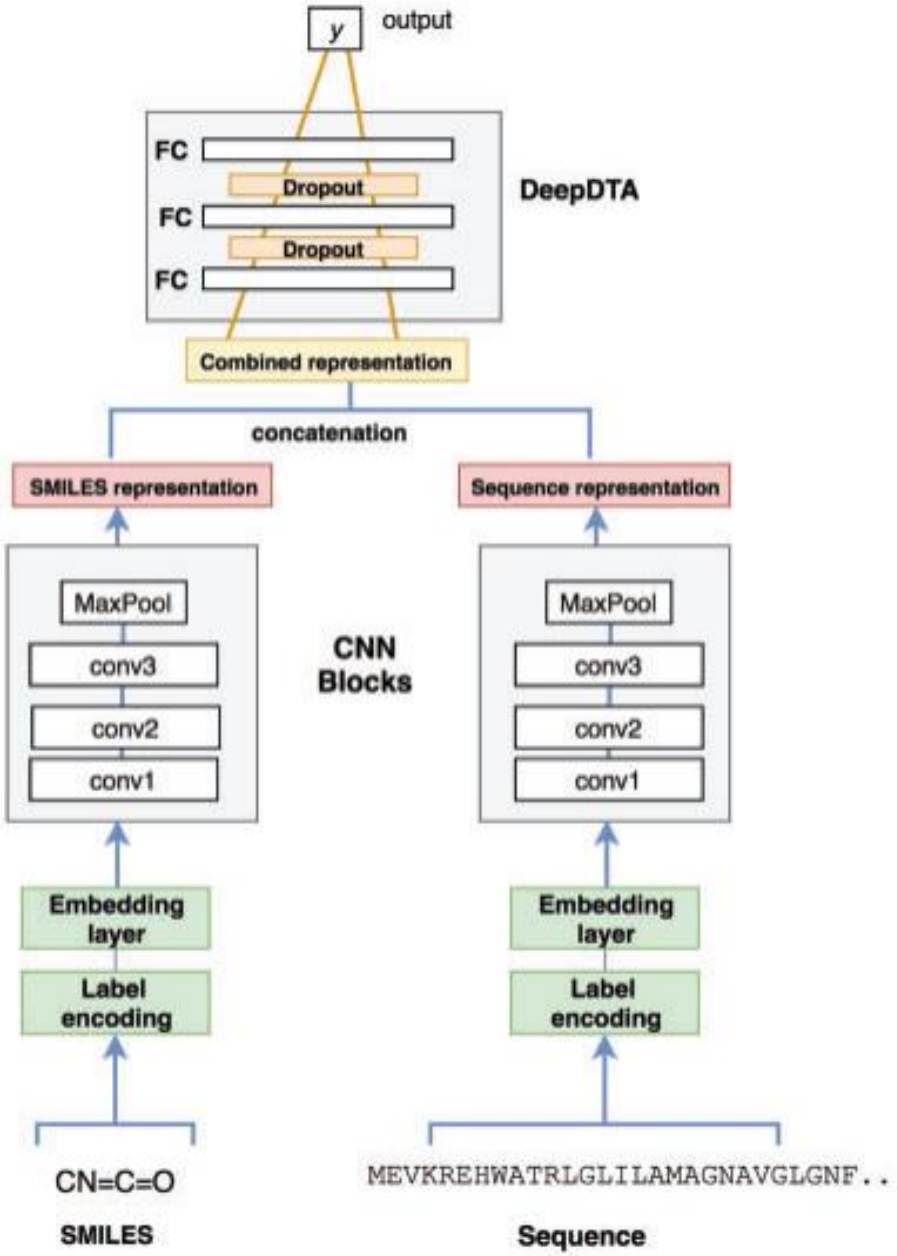
  - Drug-target pairs에 관한 Object-based features, Network-based features, Network-based features (from a heterogeneous network)의 feature들을 supervised learning 기법의 하나인 gradient boosting regression tree에 적용하여 binding affinity를 측정하게 된다. 이때, SimBoost는 gradient boosting machine model 기법이다.

- Convolutional Neural Networks (CNN) with DeepDTA

  - Drug-Target pair들에 대한 큰 dataset에 대하여 CNN을 적용하여 훈련시킴으로써 binding affinity prediction model은 sequence들에 대해 의미있는 패턴들을 추출할 수 있다. 또한, 다수의 convolution layer와 pooling layer에 대하여 계층 구조를 이루고 있기에 low-level features(drugs-drugs, target-target간의 local properties)와 high-level features(drugs-targets간의 global properties)을 동시에 추출할 수 있다

  - DeepDTA(Deep Drug-Target Affinity prediction)은 두개의 독립된 CNN을 통해 drug sequences와 target sequences에 대하여 각각 학습을 하여 예측을 하는 모델이다. 각 CNN은 3개의 1D-convolutional layer로 이루어져 있으며 이를 뒤따르는 max-pooling layer가 있다. max-pooling layer에서 나온 feature들은 마지막으로 fully connected layer에서 예측을 만들 때 사용된다.

4.3 프로세스


y
output
FC
Dropout
FC
Dropout
FC
DeepDTA
Combined representation
concatenation
SMILES representation
Sequence representation
MaxPool
conv3
conv2
conv1
CNN Blocks
MaxPool
conv3
conv2
conv1
Embedding layer
Label encoding
Embedding layer
Label encoding
CN=C=O
SMILES
MEVKREHWATRLGLILAMAGNAVGLGNF..
Sequence

5 개발 일정 및 역할 분담

5.1 개발 일정

5월			6월					7월					8월					9월	
3주	4주	5주	1주	2주	3주	4주	5주	1주	2주	3주	4주	5주	1주	2주	3주	4주	5주	1주	2주
착수보고서																			
	머신러닝 관련 공부																		
						모델 기법 연구													
								Dataset 라벨링											
								여러 기법들 비교											
								여러 Dataset 실험											
										최적화된 모델 구현									
													Regularization & overfitting 줄이기						
																최종 발표/보고서 준비			

5.2 역할 분담

이름	역할 분담
박한얼	● KronRLS 모델 구현 ● 데이터 전처리 구현
김선아	● SimBoost 기법을 이용한 모델 구현 ● 데이터 전처리 구현
김연후	● CNN을 이용한 DeepDTA 구현 ● 결과물에 대한 시연을 웹사이트 상에서 구현
공통	● 보고서 작성 및 발표 ● 모델 테스트 및 성능 평가 ● 데이터 라벨링