약물 디자인을 위한 Residual LSTM

유소영, 김정현*
순천향대학교, 세종대학교*
yooso0731@sch.ac.kr, j.kim@sejong.ac.kr*

Residual LSTM for De Novo drug design

Soyoung Yoo, Junghyun Kim*
Soonchunhyang Univ., Sejong Univ.*

요 약

RNN 기반 모델을 사용한 새로운 분자식(SMILES) 생성은 가능성 있는 후보약물을 발견하도록 도와 신약 개발에 걸리는 시간 및 비용을 효과적으로 줄일 수 있다. 분자 생성 성능 최적화를 위한 여러 실험이 선행 연구에서 진행되었지만 제안된 모델의 복잡도가 높았다. 모델의 높은 복잡도는 학습 소요 시간의 증가로 이어지기 때문에 우리는 LSTM의 유닛 수를 적절히 조정하여 기존 모델보다 복잡도를 낮춘 Residual LSTM을 제안한다. 실험 결과, 제안 모델이 기존 모델보다 6배 이상 복잡도가 낮았고 생성 성능은 유효성에서 3% 높은 95%를 보였다.

Ⅰ. 서 론

약물은 후보 물질 탐색 및 개발, 일련의 임상 시험 단계를 거쳐 시판되므로 그 과정에 많은 시간과 비용을 필요로 한다.[1] 연구에 따르면 5000개의 후보 약물 중에서 5개만이 전임상 시험을 통과하고 그 중 단 하나가 임상 시험에 통과하여 시장에 출시된다.[2] 따라서 가능성 있는 후보약물들을 발견하는 것은 신약 개발 전체 과정에서의 시간 및 비용 최소화를 위해 중요하다. 최근에는 컴퓨터 자원을 사용하여 후보 물질을 생성하거나 식별하는 연구가 많이 발표되고 있다.

약물 디자인(drug design)은 생물학적 표적에 작용할 수 있는 새로운 약물을 개발하는 과정이다. 이때 인공지능은 방대한 화학 공간에서 아직 알려지지 않은 약물의 발견을 돕기 위해 사용되며 주로 RNN(Recurrent neural network) 기반 모델로 새로운 약물의 분자식을 생성한다.[3] Simplified molecular-input line-entry system(SMILES)은 화합물의 분자식을 나타내는 선형 표기법으로, 원자에 해당하는 일련의 문자와 연결성을 나타내는 특수 문자로 구성된다.[4]

Santos, B. P 등은 유효한 분자 생성을 보장하는 순환 네트워크를 위한 여러 성능 비교 실험을 진행했다.[5] 저자들은 여러 하이퍼 파라미터가 생성된 분자들의 유효성과 다양성에 미치는 영향을 확인했다. 그 결과 모델 학습에 사용되는 분자 데이터 셋의 크기는 클수록, 배치 크기는 작을수록 유효한 분자식을 많이 생성했다. 하지만 그만큼 학습 소요 시간도 증가시키므로 연구자의 적절한 선택이 필요하다. Sampling temperature는 0.75로 설정되었을 때 유효성과 다양성 중 한 쪽에만 치우쳐지지 않는 분자식을 생성했다. 모델은 그림1-(a)와 같이 2-계층 LSTM으로 구성했을 때 성능이 좋았다. 하지만 해당 실험에서 모델의 복잡도는 고려되지 않았다. 복잡도가 높은 모델의 학습 소요 시간은 학습 데이터 셋의 크기는 커질수록, 배치 사이즈 수는 줄어들수록 가중된다.

Residual LSTM[6]은 ResNet[7]에 영감을 받아서 기본 LSTM에 잔차 연결(Residual connection) 과정을 도입한 모델이다. 학습 모델이 너무 많은 층으로 구성되어 있거나 뉴런의 수가 클 경우 입력 정보는 여러 층을 거치면서 정보 손실이 발생할 수 있고 가중치 학습이 잘못되는 문제가 발생할 수도 있다. 따라서 잔차 연결은 이전 층의 정보를 연결하여 이러한 문제를 해결한다. 잔차 연결 과정은 모델의 학습 난이도를 낮춰 모델의 빠른 수렴을 도우며 결과적으로 모델의 성능을 향상시킨다.

본 논문에서는 이러한 잔차 연결을 사용하여 선행 연구 모델보다 복잡도는 낮지만 좋은 생성 성능을 보이는 모델을 제안하고자 한다.

Ⅱ. 본론

우선 우리는 잔차 연결을 통한 성능 개선 효과를 확인하고자 선행 연구 모델에서 잔차 연결 유무에 따른 생성 성능을 비교했다.(그림 1) SMILES 문자열은 토큰화와 원-핫 인코딩을 거친 후 모델에 입력된다. LSTM의 유닛 개수는 512이고, Dropout 비율은 0.2이다. 이후 Dense 계층을 통해 모든 단어에 대한 등장 확률이 계산되며 잔차 연결이 있는 모델의 경우 입력 계층과 연결(잔차 연결)된다. 마지막으로 Temperature sampling 기법으로 다음 시점의 SMILES 문자를 생성한다.

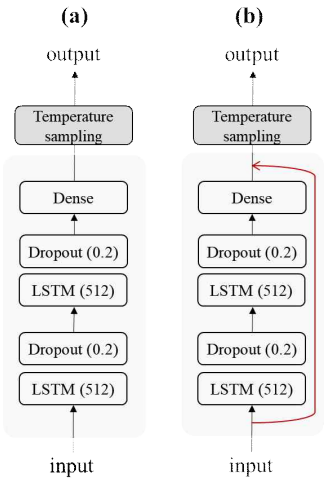

그림 1 (a) 선행 모델,
(b) 선행 모델+잔차 연결

모델 학습 시에는 잘못된 예측으로 인한 오류의 누적을 방지하기 위해 Teacher forcing 알고리즘이 사용되고 학습이 완료된 모델은 현재 시점에서 예측된 값이 다음 시점에 입력되어 누적된 분자식을 생성한다. 성능 검증에는 생성된 분자 셋의 유효성과 고유성이 사용된다. 이때 유효성은 생성된 분자 셋에서 유효하게 구성되는 분자의 비율(%)이고, 고유성은 생성된 분자 셋에서 중복되지 않은 분자의 비율(%)이다.

본 연구에서 모델 학습에는 선행 연구[5]가 진행될 때 수집된 ChEMBL 데이터베이스[8]의 데이터 중 임의로 추출된 100,000 SMILES 문자열이 사용됐다. 이 데이터는 모두 RDKit에 의해 필터링되고 정규화됐다. 손실함수는 Categorical cross-entropy이며 다른 하이퍼 파라미터는 선행 연구의 결과를 참고해 구성하였으므로 최적화 알고리즘은 RMSprop, Temperature는 0.75이다.

각 모델을 배치 크기 16에서 16번 반복 학습한 후 새로운 분자를 100개 생성하여 유효성과 고유성을 확인했다. 두 모델의 학습 파라미터 수는 3,234,337개였고 유효성과 고유성은 순서대로 그림1-(a)모델의 경우 92%, 100%, 그림1-(b)모델의 경우 97%, 100%였다. 그림1-(b)모델은 그림1-(a)와 복잡도가 동일했지만 유효성은 5% 정도 개선된 결과를 통해 잔차 연결의 성능 개선 효과를 확인할 수 있었다.

최종적으로 본 논문에서 제안하는 분자식 생성 모델의 구조는 그림 2와 같다. 모델의 복잡도 감소를 위해 LSTM의 유닛 개수는 순서대로 256, 128으로 변경하고 성능 향상을 위해 잔차 연결 과정을 추가했다.

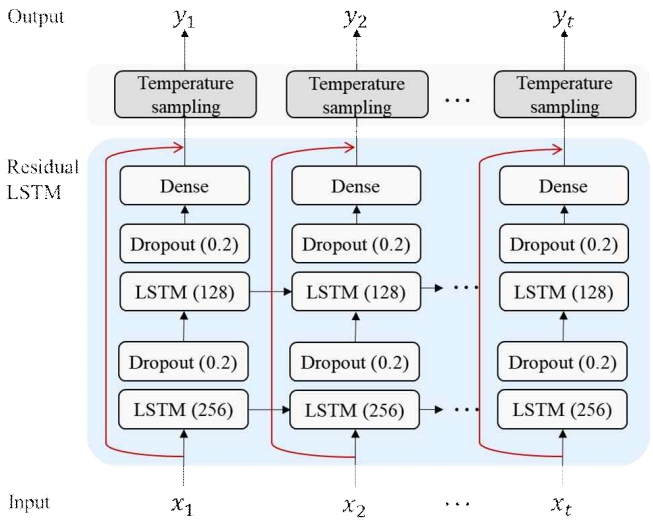

그림 2 제안 모델 구조

제안 모델 또한 배치 크기 16에서 16번 반복 학습한 후 새로운 분자를 100개 생성하여 유효성과 고유성을 확인했다. 그 결과 순서대로 95%, 100%의 성능을 보였다. 표 1은 제안 모델과 선행 연구[5, 그림1-(a)] 모델의 성능 비교 표이다. 이를 통해 제안 모델의 학습 파라미터 개수가 비교 모델에 비해 6배 이상 적음에도 생성 분자 셋의 유효성은 3% 향상된 것을 확인할 수 있다.

Model	Valid %	Unique %	# of trainable parameters	Time
Ref.[5]	92	100	3,234,337	45:40
Prop.	95	100	498,337	18:40

표 1 제안 모델과 선행 연구의 모델 성능 비교

Ⅲ. 결론

우리는 후보약물 발견을 위해 유효한 분자식을 다양하게 생성하는 Residual LSTM을 제안한다. 선행 연구에서 생성 성능 최적화를 위한 여러 비교 실험이 진행되어 하이퍼 파라미터에 따른 영향을 확인할 수 있었지만 모델의 복잡도는 고려되지 않았다. 모델의 높은 복잡도는 학습 소요 시간의 증가로 이어지며 학습 데이터 셋의 크기가 증가할수록 소요 시간은 가중된다. 따라서 우리는 기존 모델보다 복잡도가 낮지만 분자 생성에 효율적인 모델을 제안한다. 실험 결과, 제안 모델은 기존 모델에 비해 복잡도가 6배 이상 낮았지만 유효성은 3% 향상된 95%임을 확인할 수 있었다.

잔차 연결은 깊게 구성된 모델의 기울기 소실 혹은 발산의 문제를 완화할 수 있으므로 깊게 구성된 모델에서 성능이 더 좋아질 수 있다. 또한 모델의 생성 성능은 샘플링 기법에 따라 달라질 수 있다. 따라서 향후에는 더 깊은 모델에 대한 실험과 adaptive temperature sampling 기법, Top-P 기법 등 여러 샘플링 기법에 따른 성능 비교 실험을 진행할 예정이다.

참 고 문 헌

[1] Kola, I., and Landis, J. “Can the pharmaceutical industry reduce attrition rates?,” Nature reviews Drug discovery, 3.8, pp. 711-716, 2004.

[2] Torjesen, I. "Drug development: the journey of a medicine from lab to shelf." Pharmaceutical Journal, 2015.

[3] Mouchlis, Varnavas D., et al. "Advances in de novo drug design: From conventional to machine learning methods." International journal of molecular sciences 22.4: 1676, 2021..

[4] Weininger, D. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." Journal of chemical information and computer sciences 28.1, pp. 31-36, 1988.

[5] Santos, B. P., et al. "Optimizing Recurrent Neural Network Architectures for De Novo Drug Design." 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2021.

[6] Huang, L., Xu, J., Sun, J., & Yang, Y. "An improved residual LSTM architecture for acoustic modeling." 2017 2nd International Conference on Computer and Communication Systems (ICCCS). IEEE, 2017.

[7] He, K., Zhang, X., Ren, S., & Sun, J. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[8] Gaulton, A., et al. "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic acids research 40.D1, D1100-D1107, 2012.