논문 요약 보고서

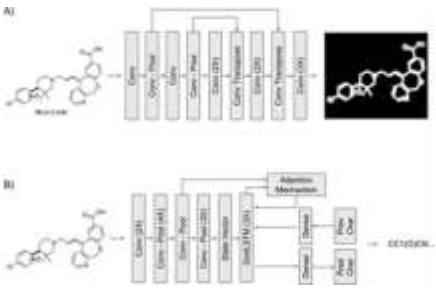
논문 제목: Molecular Structure Extraction from Documents Using Deep Learning
저자명: Joshua Staker, Kyle Marshall, Robert Abel, and Carolyn M. McQuaw
출판 정보: J. Chem. Inf. Model. 2019, 59, 3, 1017−1029

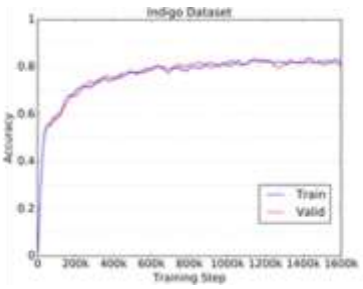
- 소개 배경

분자와 관련된 데이터를 포함하는 대부분의 간행물은 분자 구조를 컴퓨터가 읽을 수 있는 형식으로 제공하지 않기 때문에 Drug discovery에서 대량의 실험 데이터를 생성하는 데에는 어려움이 따른다. 개발의 가속화를 위해서는 데이터 추출 및 분류 과정을 최대한 자동화하는 것이 필요하다. 이에 대한 해결책으로는 기존에 소개된 여러 방법들이 있지만, 저해상도 이미지에 대한 output의 quality의 변동과 같은 난제에 봉착한다. 본 논문의 목표는 문서에서 바로 SMILES를 추출할 수 있는 추출 방법을 개발하는 것, 그리고 이러한 시스템을 이용하여 저화질 이미지의 예측 정확도 또한 향상시키는 딥러닝 모델의 실현을 보여주는 것이다.

- 결과

문서에서 화학 구조를 추출하고, 이미지에 대한 SMILES 표현 예측을 하기 위한 딥러닝 모델 솔루션을 제시하는 데에 성공했다. 이 방법을 통해 어떤 이미지인지에 관련 없이 (저화질 이미지 포함) 문서 내의 데이터에 대한 학습 및 예측을 진행할 수 있다.





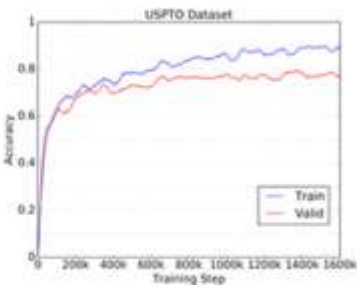


Table 1. Overall Performance against Validation and Test Sets

data set	accuracy
Indigo validation set	82%
USPTO validation set	77%
Valko data set	41%
proprietary data set	83%

여러 문서 Dataset에서 추출한 분자 이미지를 사용하여, 우리의 모델이 상당히 높은 정확도로 문서에서 분자의 이미지를 예측하는 것을 알 수 있었다. 정확도는 다음(Table 1)과 같다.

- 시사점

본 논문에서 소개한 모델을 사용한다면, 저해상도 이미지의 분할과 예측 모두 좋은 성과를 거둘 수 있다. 따라서 본 화학 구조 추출 알고리즘의 개발 및 활용은, 여러 면에서 Drug Discovery분야에서의 개발 속도를 크게 가속화할 수 있을 것으로 예상한다.