

분자구조 이미지 SMILES 변환 Hackathon

Eunhui Kim/KISTI

목차

분자구조 이미지 SMILES 변환 AI경진대회

- 1) 화학 분자 이미지 텍스트 변환 SMILES?
- 2) 3위 UNIST 랩 요정들의 이야기
- 3) 데이터 생성 및 샘플링
- 4) 데이터 전처리
- 5) 모델 학습
- 6) Ensemble ?!
- 7) GPU 멀티 프로세싱

1) 화학 분자 이미지 텍스트 변환 SMILES?

**LG AI
HACKATHON**
분자구조 이미지 SMILES 변환

대회 기간
2020년 9월 1일 ~ 10월 9일

수상 혜택
총 1,000만원 상당의 상금과 채용 우대

참가 방법
<https://dacon.io>

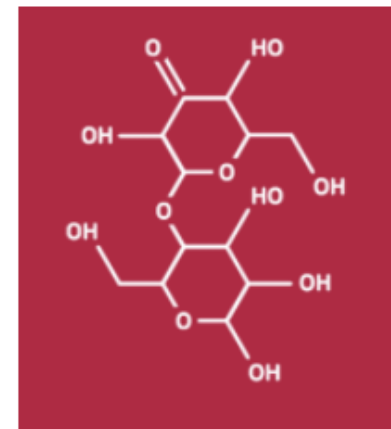
  **LG**
 **DACON**





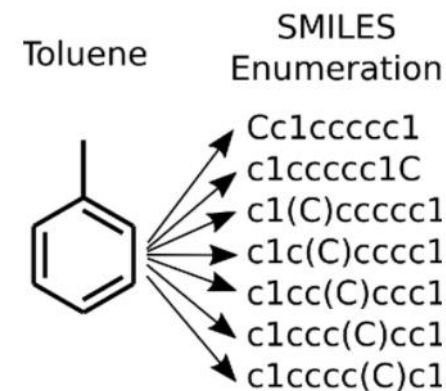
1) 화학 분자 이미지 텍스트 변환 SMILES?

- 물 분자는 수소 원자 H 2개와 산소원자 1개로 구성됨. 이를 H₂O라는 분자로 표현하는 데 이를 분자식 이라 함.
- SMILES(Simplified Molecular Input Line Entry System)은 기존의 분자식 보다 컴퓨터에서 다루기 편한 표현 방법으로, 최근 신약 연구 및 약물 재창출 연구에서 자주 사용됨.



- 단일 분자 구조에 대해 여러 개의 SMILES 표현식이 존재함. 따라서 모델의 성능 평가를 정확도(accuracy)가 아닌 정답지의 SMILES표현식이 얼마나 유사한지를 측정하는 Tanimoto Similarity가 사용됨.

Tanimoto Similarity는 IoU(Intersection over Union)개념으로 두 표현식 사이의 교집합의 크기를 합집합의 크기로 나누어 계산됨.



2) Hacthathon 3위 UNIST 랩 요정들의 이야기

분자구조 이미지 SMILES 변환 AI 경진대회

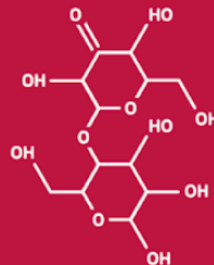
산업 | LG | 화학분자 이미지 텍스트 변환 SMILES | Similarity

상금 : 1,000만원

2020.09.01 ~ 2020.10.09 17:59




[+ Google Calendar](#)

547팀 마감



참여

● WINNER ● 1% ● 4% ● 10%

#	팀	팀 멤버	최종점수	제출수	등록일
1	PBRH		0.99861	-	-
2	bbchip13		0.9979	-	-
3	랩요정들		0.99567	-	-



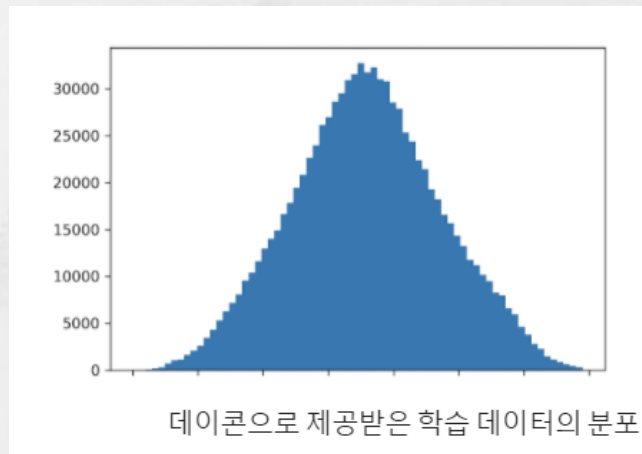
3) 데이터 생성 및 샘플링

[주어진 데이터의 한계 파악]

- 대회에서 제공한 학습 데이터에는 표현식의 길이가 70자를 넘어가는 데이터는 포함되지 않음.
- 학습 데이터만으로는 더욱 길고 복잡한 표현식을 포함하는 테스트 데이터를 제대로 맞추는 것이 어렵다고 판단.

[대안]

- Pubchem 에서 제공하는 다양한 길이의 SMILES 데이터를 활용하여 학습을 위한 300 x 300 크기의 분자구조 이미지를 추가적으로 1억 개 정도 생성.
- 데이터의 분포를 확인한 결과, 표현식의 길이가 100 이하인 데이터의 비중은 약 95% (106,595,711개).
- 최종적으로는 길이가 100 이하인 데이터만을 집중해서 생성



3) 데이터 생성 및 샘플링

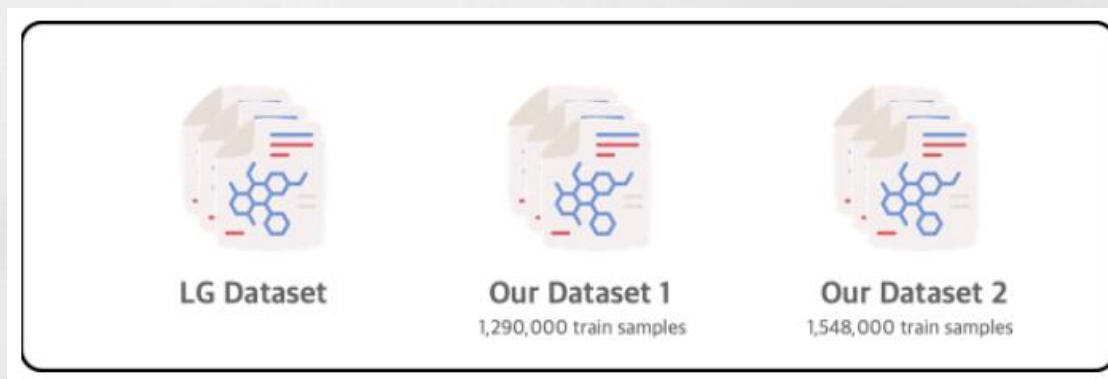
[데이터 학습을 위한 데이터 량의 한계]

- 길이가 100 이하인 데이터만 생성하였으나, 대량의 데이터인 관계로 모두 학습에 사용하는 것이 불가능함.



[대안]

- Uniform Distribution을 따르도록 각 표현식 길이 별로 35,000개의 데이터를 랜덤 샘플링 함.
- 학습 시간을 고려하여 앞선 분포에서 다시 한번 랜덤 샘플링을 진행함.
- 대회에서 제공받은 데이터 셋과 대안 방식으로 샘플링 한 데이터 셋 2개를 모델 학습에 사용함.



4) 데이터 전처리

- 일반적인 표현식 모델에 맞춰 모델에 바로 들어갈 수 있도록 전처리
- ImageNet에서 사전 학습된 (pre-trained) 모델을 사용하고자, ImageNet Setting에 맞게 이미지 Resize 및 Regularization을 적용.
- 표현식에 대해 SMILES 데이터셋에서 나올만한 Token을 모두 고려하여 Token Dictionary를 구성함.
- Image파일은 hdf5 파일로, token 파일은 json 파일로 저장해서 학습 과정 중 바로 호출 가능하도록 처리 함.

5) 모델 학습

- 모델로는 이미지에 captioning의 대표적인 방법 중 하나인 2015년 CVPR에 발표된 Google의 Show and Tell 논문 사용

Show and Tell: A Neural Image Caption Generator

4089회 인용

Oriol Vinyals
Google

vinyals@google.com

Alexander Toshev
Google

toshev@google.com

Samy Bengio
Google

bengio@google.com

Dumitru Erhan
Google

dumitru@google.com

Abstract

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify

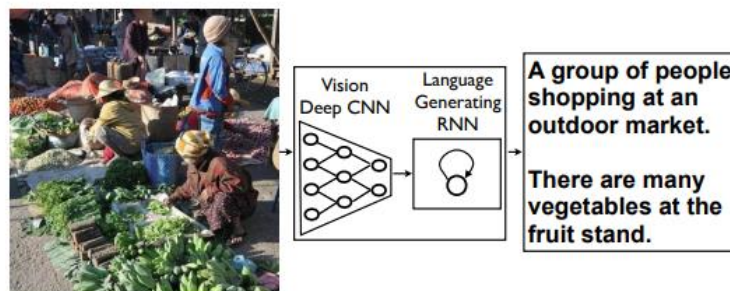
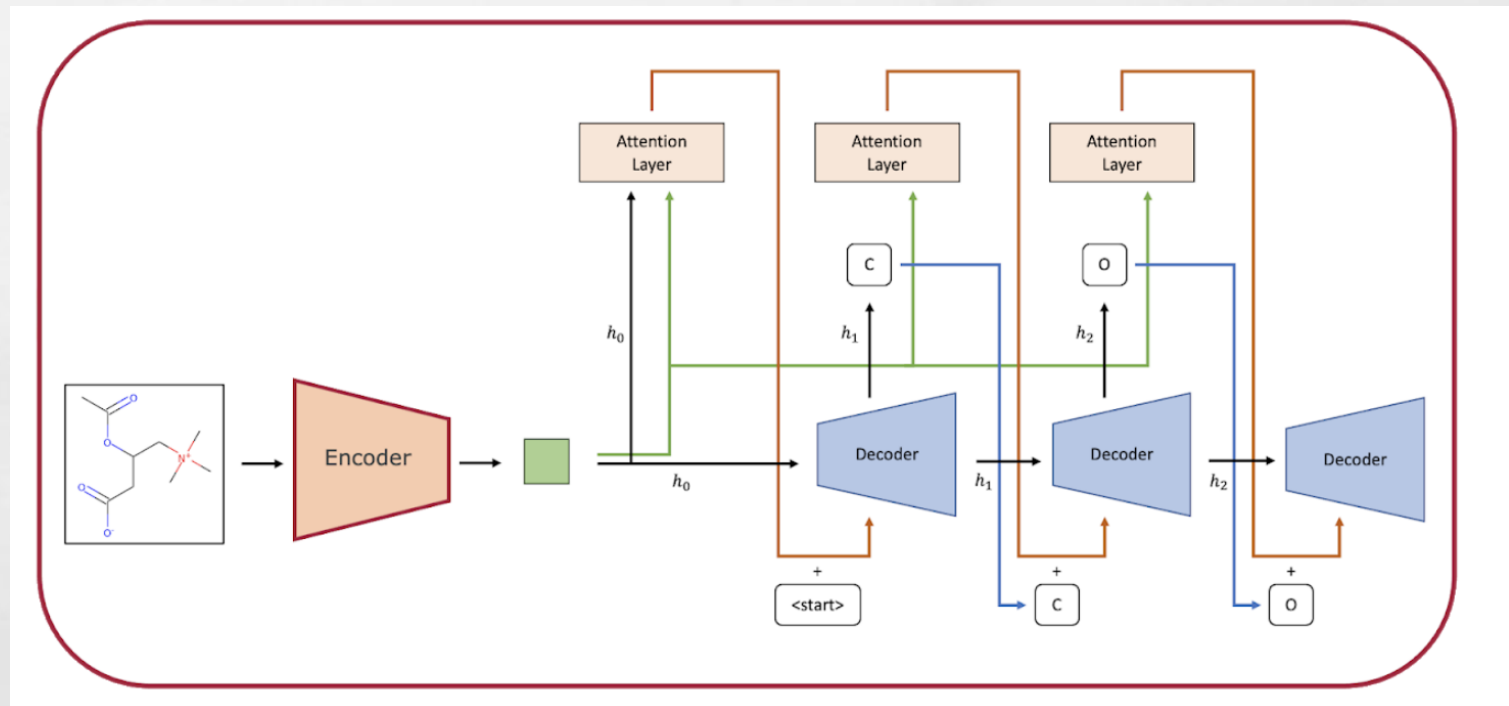


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

5) 모델 학습

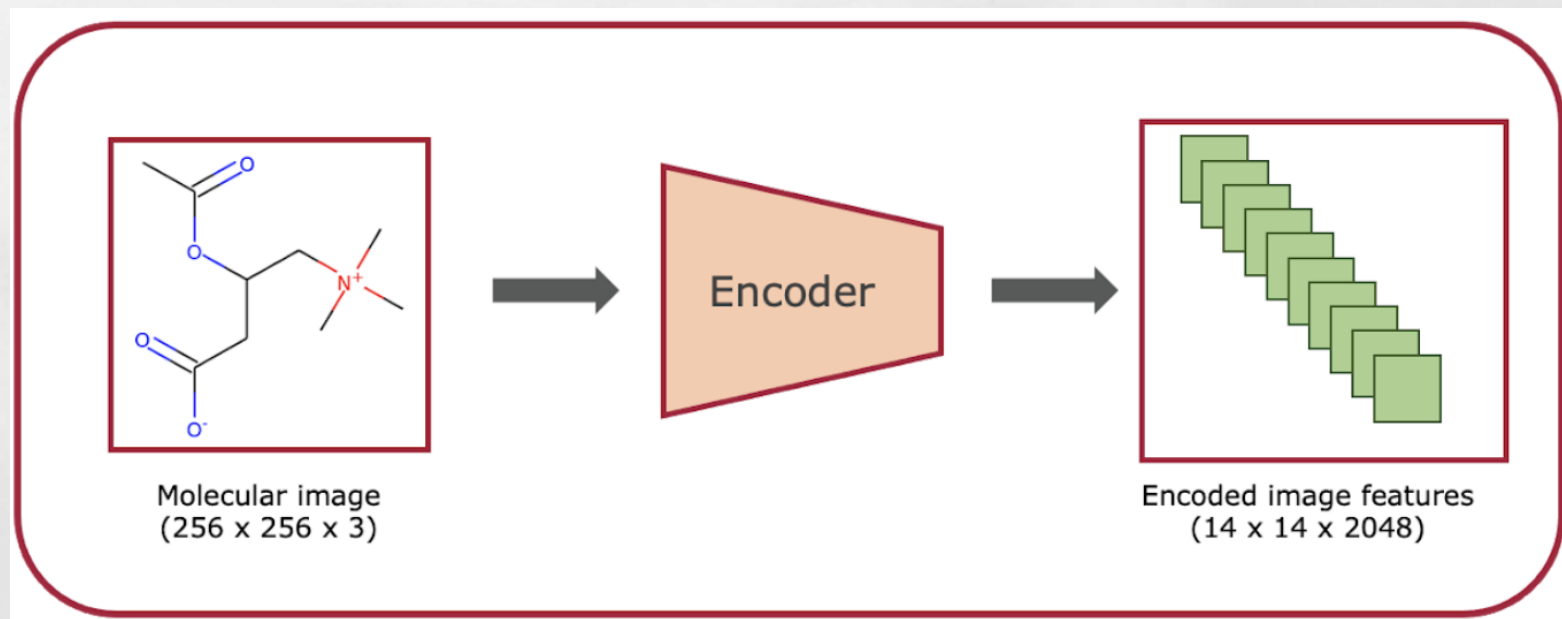
- 모델의 구성은 Encoder, Attention, Decoder 세 부분으로 나뉨
 - Encoder : 이미지를 인식하고 해당 이미지에서 필요한 특징(feature)를 추출
 - Attention : 학습 과정에서 이미지의 어떤 부분에 집중할 지를 인지하는 부분
 - Decoder : Encoder와 Attention의 정보를 바탕으로 Token을 순차적으로 생성



5) 모델 학습

▪ Encoder Network

- Encoder로는 기 제시된 모델들인 EfficientNet-B0, ResNeXt-101-32x8d, WideResNet-101-2, ResNet-152 네 모델을 사용함.
- EfficientNet 계열에 비해 ResNeXT계열이 성능이 좋게 나옴
- ImageNet pre-trained 된 모델에서 전 5개 layer를 제외한 나머지 layer parameter들을 분자구조 이미지를 추가로 학습 시켜 업데이트 하여 Fine-Tuning 함.



5) 모델 학습

▪ Attention Network

- Show, Attention and Tell 모델의 Attention 네트워크 활용
- 분자구조 이미지 중 각 원소마다 결합 구조를 풀어나가는 Attention 정보를 활용하고자 함



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

JMLR 2015, 6073회 인용

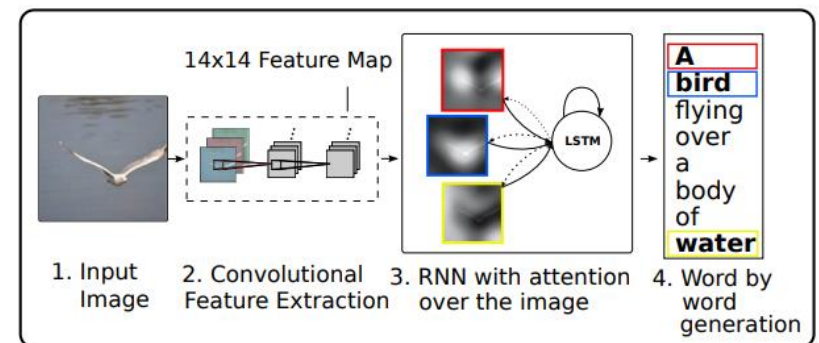
Kelvin Xu

KELVIN.XU@UMONTREAL.CA

Abstract

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. We

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4



5) 모델 학습

▪ Attention Network

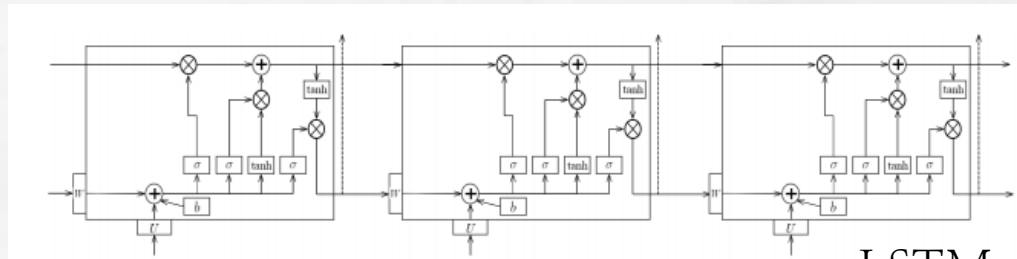
- Attention 네트워크는 두 입력을 받는데,
하나는 Encoder에서 나온 이미지 feature vector 이고,
다른 하나는 Decoder의 hidden 벡터임.
- 해당 벡터들을 미리 설정한 Attention 차원에 맞춰주고,
두 벡터들을 합침.
- 이미지의 크기에 맞춰 벡터를 줄이는 작업 후,
SoftMax Layer에 적용하여
각 token마다 이미지의 어떤 부분에 집중을 해야하는지 학습하게 함



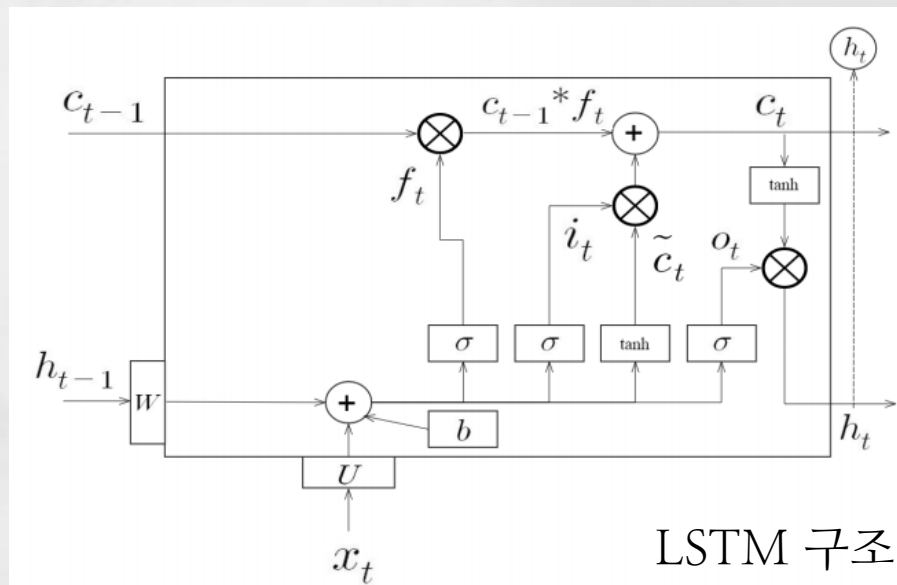
5) 모델 학습

■ Decoder Network

- LSTM Cell로 구성됨. SMILES 특성상 괄호가 열리면 뒤에서는 괄호가 닫혀야 하는데 장기 패턴을 고려하여 LSTM Cell을 사용함.
- Decoder Network의 입력은 encoding된 이미지에 Attention 네트워크의 가중치를 곱한 Vector, 그리고 이전 Hidden State의 입력을 받음.
- 출력 Hidden State는 fully connected layer를 거쳐 token에 대한 확률 값으로 표현됨.



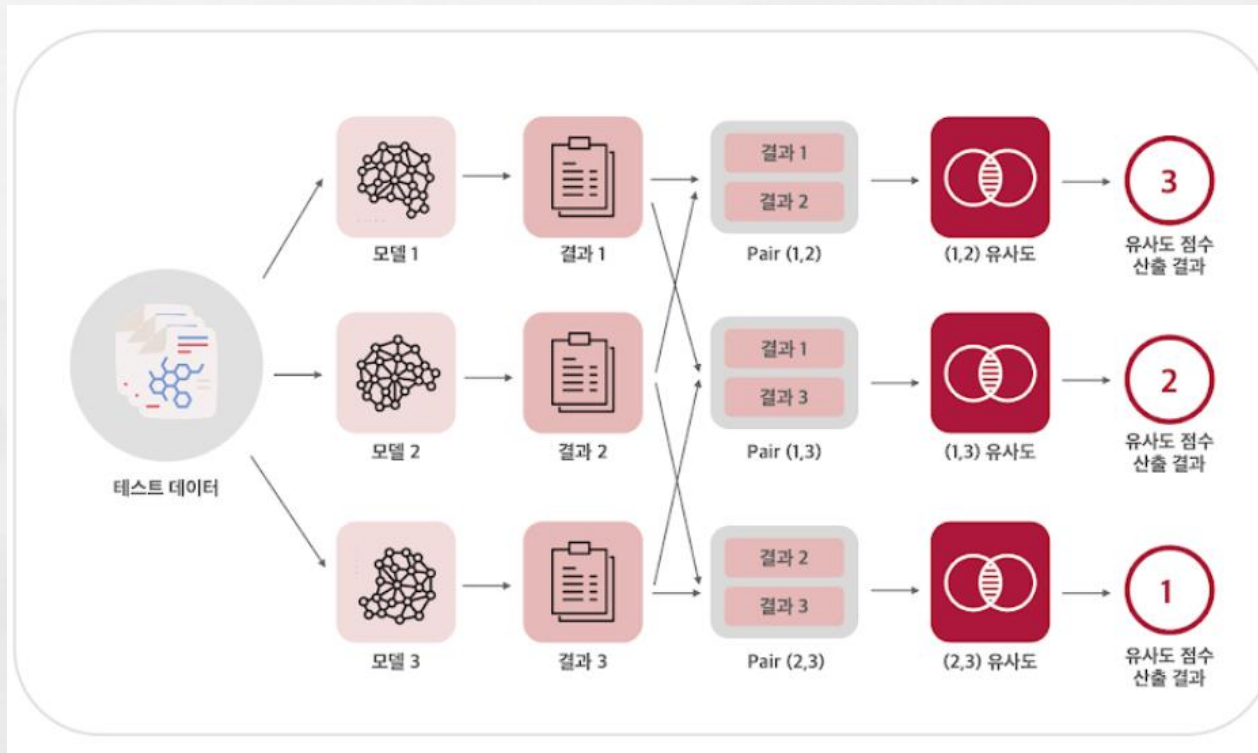
LSTM network



LSTM 구조

6) Ensemble ?!

“모델 페어 기반의 앙상블?”



고득점 모델 선택

	모델1	모델2	모델3
Pair(1,2) 유사도점수	3	3	-
Pair(1,3) 유사도점수	2	-	2
Pair(2,3) 유사도점수	-	1	1
누적 점수	5	4	3

6) Ensemble ?!



동작 속도 비교

단일 모델 평균 동작 속도 vs 앙상블 모델 동작 속도

단일 모델 0.0934 sec

앙상블 모델 0.1992 sec

최대 성능 비교

단일 모델 최대 성능 vs 앙상블 모델 최대 성능

단일 모델 0.9729

앙상블 모델 0.9961

[앙상블에 사용된 성능 top 5 단일 모델]

Model Number	Encoder Model	Pretrained	Attention/Decoder /Embedding Dimension	Data Set	Best Public Score	Prediction Time
1	Wide ResNet101-2	True	512	Our Dataset 2	0.9729	0.0853
2	Wide ResNet101-2	True	512	Our Dataset 1	0.9625	0.0861
3	ResNet152	True	512	LG Dataset	0.9622	0.0947
4	ResNet152	True	256	LG Dataset	0.9512	0.1035
5	ResNeXt-101-32x8d	True	256	Our Dataset 1	0.9677	0.0972

LG Dataset : 700,000 train samples / Our Dataset 1 : 1,290,000 train samples / Our Dataset 2 : 1,548,000 train samples

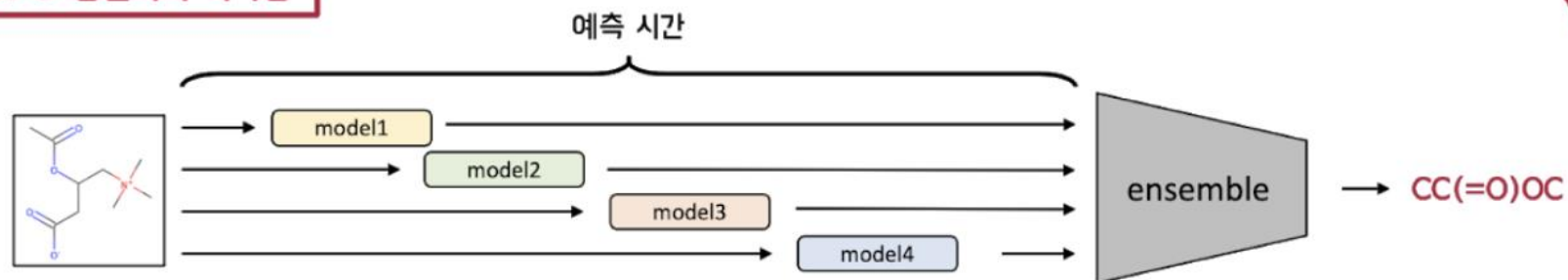
* Prediction Time의 단위는 초이며, 하나의 테스트 데이터를 1개의 GPU를 사용하여 예측하는데 소요된 시간.



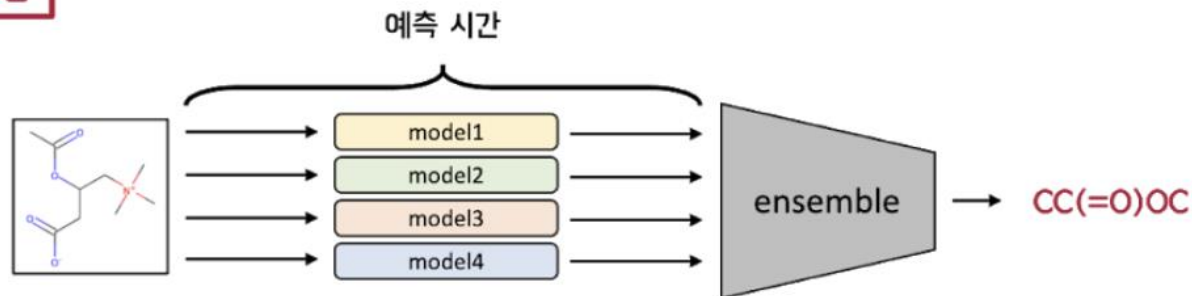
7) GPU 병렬처리



GPU 병렬처리 미사용



GPU 병렬처리 사용



감사합니다.

