transformer모델의 attention score를 활용한 개체별 유전체 데이터 핵심부분 연구

전석환, 이주원, 홍성은, 방준일, 김화종
강원대학교 컴퓨터공학 전공
tonywjs@naver.com, alcatraz76@naver.com,
sungkenh@gmail.com, tkfka965@gmail.com, hjkim3@gmail.com

A study to identify key parts of genetic data by breed using the Attention Score of Transformer model

Jeon Sukhwan, Lee Joowon, Hong Sungeun, Bang Junil, Kim Hwajong
Kangwon National Univ.

요 약

유전체 데이터는 앞으로의 의료 빅데이터의 핵심 자원으로 주목받고 있다. 유전체 데이터는 한 개체의 모든 유전정보를 뜻하며 한 개체의 유전체 데이터는 수십억 개의 염기서열로 이루어져 있어 이를 모두 분석하는 일은 간단하지 않다. 본 논문에서는 dacon의 '유전체 정보 품종 분류 AI 경진대회'의 dataset으로 Transformer 모델의 attention score를 통해 유전체 데이터의 핵심 위치를 확인하는 실험을 하였다. 확인 결과 품종별 attention score가 비교적 동일한 부분에 높은 값이 출력 됨을 확인하였다. 이후 큰 규모의 유전체 데이터를 사용한다면 해당 개체의 더욱 다양한 관련성을 찾을 수 있을것이라 예상한다.

Ⅰ. 서론

전 세계적으로 유전체 데이터는 앞으로의 의료 빅데이터 핵심 자원으로 주목받고 있다. 각 나라의 개인정보보호법에 의해 사용이 까다롭던 유전체 데이터를 비교적 최근 비식별화 기법을 통해 안전하게 수집 및 사용이 가능하도록 법이 개정되며 많은 나라에서 유전체 데이터를 활용한 연구 및 사업이 활발히 진행되고 있다. 유전체 데이터에 딥러닝 적용 시 신약개발, 질병 예측, 품종 분류 등 다양한 분야에 활용 가능하다.

유전체 데이터는 한 개체의 모든 유전정보를 뜻하며 개체 및 품종에 따라 각기 다른 변이 양상을 나타내고 있다. 이중 대략 1,000개의 염기서열마다 1개 꼴로 나타나는 SNP 데이터는 개체별 각기 다른 단일 염기 위치를 말하며 사람의 경우 약 100만개의 SNP를 갖는다.[1] 만약 생물학적 특성에 따라 SNP 정보가 달라지는 부분을 알 수 있다면 유전체 전체를 분석할 필요 없이 가장 변화가 뚜렷한 부분만을 분석할 수 있기에 시간과 비용을 절감할 수 있다.

본 논문에서는 Transformer 모델의 attention score를 통해 유전체 데이터를 사용한 품종 예측 시 어떤 부분의 유전체 정보에 가중을 두어 예측을 하였는지 확인하여 품종에 따른 유전체 정보의 핵심 위치를 확인하였다. 결과 판단 지표로는 attention score가 높은 위치가 품종별로 유사하게 위치하는지, 특정 위치에 확연히 높은 attention score가 존재하는지로 판단하였다.

Ⅱ. 본론

Transformer 모델은 2017년 구글의 'Attention is all you need' 논문[]에서 발표한 모델로 기존 자연어 처리 방식의 순환신경망(RNN)을 사용하지 않고, 기존 seq2seq의 구조인 인코더-디코더 구조에서 어텐션 메커니즘만을 사용하여 기존의 RNN, LSTM보다 높은 성능을 보여주었다.

기존 자연어 데이터 처리에서 사용되었던 순환신경망 RNN과 LSTM은 특정 길이의 단어 입력을 순차적으로 받으며 입력 순서에 따른 문장의미를 추론한다. 순차적인 input 데이터 입력과 동시에 이전까지의 단어 정보가 들어있는 hidden state를 같이 입력받으며 문장 전체가 모두 입력되었을 때 문장 전체의 의미를 가지고 있는 hidden state가 완성된다. 다만 문장의 길이가 길어질수록 입력 횟수가 많아져 hidden state 내의 초기 단어의 의미가 희석된다. 이를 RNN 기울기 소실이라고 하며 이를 보완하기 위해 LSTM이 등장했다. LSTM은 이전 입력정보인 hidden state와 현재 cell의 정보를 각각 얼마나 반영할지를 결정하는 gate를 만들어 RNN의 기울기 소실을 보완하였다. 하지만 연산량이 증가하였으며 기울기 소실문제를 완전히 해결하지 못하였다. 추가로 RNN과 LSTM 모두 순차적인 입력을 받는 순환신경망의 입력 특성상 GPU의 장점인 병렬연산이 불가능하다.

Transformer 모델은 위의 문제점을 모두 해결하였다. 순차적 입력이 아닌 한번에 문장의 전체를 입력받는 Transformer 모델은 문장의 어순정보를 받을수 없기에 positional encoding을 통해 문장 내 단어의 위치정보를 가지게 한다. 이후 문장의 전체를 입력받아 각 단어들 간 어텐션 메커니즘을 동해 어텐션 스코어 및 어텐션 값를 출력한다. 이후 position-wise FFNN을 통해 입력과 동일한 크기의 레이어를 출력한다. Transformer 모델의 인코더 블록은 positional encoding 이후부터 FFNN까지이며 아래 그림과 같다. 본 논문에서는 품종 예측을 위해 인코더 부분만 사용했으며 Transformer 디코더 부분의 설명은 생략한다.

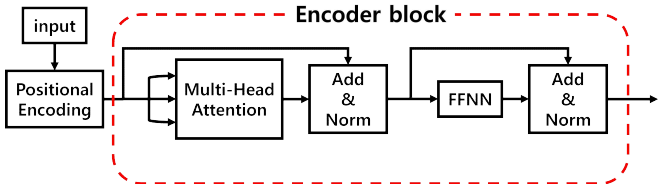
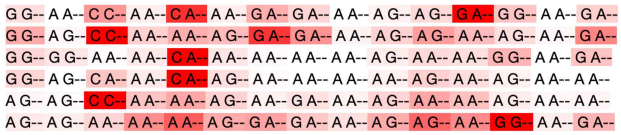


그림 1. 인코더 블록 구조

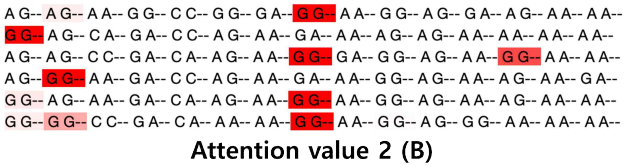
Attention value 1 (A)

Attention value 1 (B)



Attention value 1 (C)

Attention value 2 (A)



Attention value 2 (B)

Attention value 2 (C)

실험에 사용한 dataset은 dacon의 '유전체 정보 품종 분류 AI 경진대회'에서 제공한 유전체 SNP정보 데이터를 사용하였다. 데이터의 개수는 총 262개이며 class는 A, B, C 3개이다. 하나의 유전체 정보는 15개의 SNP정보와 father, mother, gender, trait 4가지 추가 정보로 이루어져 있다. father와 mother는 각각 부계, 모계의 고유 번호(0:Unknown)이며 gender는 개체의 성별(0:Unknown, 1:female, 2:male), trait은 개체의 표현형 정보이다. 15개의 SNP 정보는 총 6가지(AA, AG, CA, CC, GA, GG)정보로 구성되어 있다. 실험에는 15개의 SNP 정보만을 사용하였다.

실험에 사용한 모델의 전체 구조는 아래 그림과 같다. 모델의 input은 6가지 SNP정보를 one-hot encoding하여 입력하였다. 15개의 SNP 데이터의 위치정보를 위해 positional encoding을 진행하였다. 이후 인코더 블록을 두 번 거치게 하였으며 단일 어텐션만을 사용하여도 유의미한 어텐션 값이 출력되기에 기존 모델의 Multi-Head Attention을 단일 어텐션으로 변경하였다.

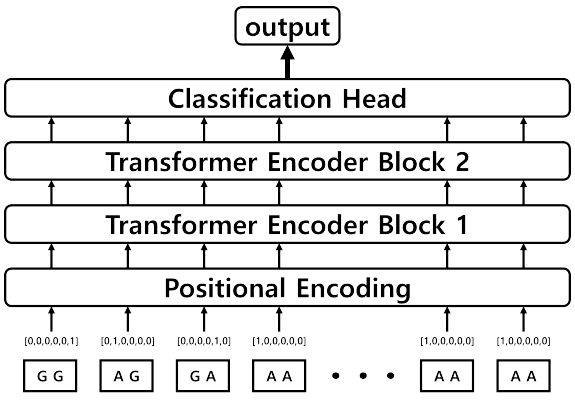


그림 3. 실험 모델 구조

epochs는 200, learning ratesms 0.0005로 학습을 진행하였으며 학습 결과는 테스트 데이터 최대 accuracy 100%, 최소 loss 0.0609로 학습이 매우 잘되었음을 볼 수 있다. 학습된 모델에 테스트 데이터셋을 입력하여 각각의 입력에 계산된 어텐션 가중치를 확인하였다. 각 15개의 SNP 정보마다 attention score를 가지며 score가 클수록 예측 결과에 기여도가 높다는 의미이다. attention score의 시각적 확인을 위해 HTML모듈을 사용하여 attention score가 높을수록 입력데이터의 색이 진한 빨간색이 되도록 하였다. 출력 결과는 위의 그림과 같다.

실험에 사용한 모델은 Multi-Head Attention대신 단일 어텐션을 사용하였으며 Transformer 인코더 블록을 두 번 사용하였기에 총 두 개의 attention score가 나온다. 첫 번째 인코더 블록에서 나온 attention score는 비교적 넓은 분포로 출력되었지만 예측 label에 따라 일관된 위치에 더 높은 가중을 두고 있음을 볼 수 있다. 두 번째 인코더 블록의 attention score는 확실한 위치에 가중이 높음을 확인하였으며 약간의 편차는 있지만 비교적 동일한 부분에 가중이 높음또한 확인하였다.

Ⅲ. 결론

본 논문에서는 유전체 SNP 데이터의 품종에 관련된 핵심부분을 transformer모델의 attetion socre을 사용하여 확인하였다. attention scroe확인 결과 class별로 비교적 동일한 위치에 attention score가 높음을 확인할 수 있었다.

논문의 실험에서는 품종 예측을 할 수 있도록 학습하였지만 품종 이외로 개체의 특성을 학습하거나 질병의 유무를 학습한 후 attention score가 높은 위치를 확인한다면 해당 label에 관련된 유전체 데이터의 핵심 위치를 확인할 수 있을 것이다.

실험에서 사용한 데이터셋은 양이 적고 유전체 데이터의 길이가 짧은 데이터임에도 유의미한 결과를 보여주었다. 이후 큰 규모의 유전체 데이터를 사용한다면 해당 개체의 더욱 다양한 관련성을 찾을 수 있을 것이다.

ACKNOWLEDGMENT

본 과제(결과물)는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.(2022RIS-005)

참 고 문 헌

[1] 김신윤, and 김태호. "유전체 다형성과 연관성 연구." 대한정형외과연구학회지 13권 1호. 7-15. 2010.

[2] ASWANI, Ashish, et al. Attention is all you need. In Advances in Neural Information Processing Systems (NIPS). 2017.