발표 대본

안녕하십니까. 한옥영 팬클럽팀인 정준영, 전은지, 오종현, 홍민서 중 발표를 맡게된 정준영입니다.

(페이지 전환- 2페이지)

저희 팀은 러닝페어 주제를 찾던 도중 영화 한 장면에서 인공지능이 논문들을 읽고 요약해주는 장면에 감명을 받아 트랜스포머 아키텍처를 활용한 영문 요약 인공지능을 개발하게 되었습니다. 정보의 바다가 되어버린 정보화사회에서 원하는 정보를 찾기가 어렵고 특히 영문 정보들은 해석과 습득에 많은 시간이 소요되기에 긴 영어 지문을 간단히 요약해주는 도구가 있다면 더 많은 정보를 빠르게 수집할 수 있을 것이라는 기대에 이 프로젝트를 진행하게 되었습니다.

(3페이지)

트랜스포머 아키텍처는 2017 nips attention is all you need 논문에서 처음으로 제안된 자연어처리 알고리즘입니다. 최근 대두되고 있는 gpt3 그리고 과거의 gpt1세대와 2세대 모두, 다시 말해 인공지능 모델의 최전선에 있는 고성능 모델들 전부 트랜스포머 아키텍쳐를 바탕으로 고안되었습니다. 기존의 시퀀스 투 시퀀스 구조를 벗어나 어텐션만을 사용했다는 것이 이전 모델들과의 차이점이라고 할 수 있겠습니다. 저희의 코드에는 총 3종류의 어텐션이 등장하는데 이에 프로젝트명을 뉴진스의 어텐션 가사 중 attention is what I want를 따서 짓게되었습니다.

(4페이지)

트랜스포머의 구조에 대해 알아보겠습니다. 전체 대략적이고 구조적인 모델 작동 방식을 먼저 살펴본 뒤에 각각의 구조인 인코더와 디코더 내부에 대해 발표하도록 하겠습니다. 트랜스포머는 크게 인코더와 디코더 부분으로 나뉘어집니다. 인코더 인풋에 원하는 데이터를 입력합니다. 파워포인트 예시에 저희는 케이크를 좋아한다는 내용의 긴 문장을 대입하도록 하겠습니다. 이후 인코더를 거쳐 나온 전체 문장의 정보가 디코더 인풋으로 대입됩니다. 디코더를 거쳐 다음 단어를 예측하고 이를 디코더에 재대입하여 반복시키면 저희가 원하고자 하는 요약문이 도출됩니다.

(5페이지)

기존 시퀀스 투 시퀀스 모델과 달리 트랜스포머는 단어를 순차적으로 입력받아 처리하는 특성이 아니기에 포지셔널 인코딩을 통해 각 단어에 벡터의 크기를 부여합니다. 이때, 각 단어가 문장의 몇번째 위치에 있는지에 따라 구분이 달라질 수 있으므로 위치 정보 역시 벡터 값에 포함시킵니다. 이후 인코더 셀프 어텐션에서 q벡터, k벡터, v벡터의 값의 가중치 값을 계산합니다. 이 가중치 계산 값은 ‘attention is all you need’ 논문의 계산식을 인용하였습니다. 스케일드 닷 프로덕트 어텐션에서 3개의 벡터를 조합하여 가중치를 계산합니다. 이를 멀티 헤드 어텐션에서 반복하였습니다. 이러한 과정에서 패딩 마스크를 통해 the, I, too 등 pad 토큰이 반영되지 않은 단어들을 매우 작은 음수값을 곱함으로써 없는 것처럼 처리하였습니다. 그 다음으로 피드 포워드 레이어에는 포지션 와이즈 피드 포워드 방식을 사용했습니다. 앞선 멀티 헤드 여텐션의 엔드값으로 도출된 가중치 행렬 w1(dmodel, dff)과 가중치 행렬 w2(dff, dmodel)의 연산을 담당합니다. 포지션 와이즈 피드 포워드의 가중치 행렬 계산식 역시 논문의 FFNN 수식을 참고하였습니다. 마지막으로 정규화 과정을 거치는데 이는 텐서의 마지막 차원에 대해 평균과 분산을 구해 정규화시켜 학습을 돕는데 저희는 LayerNormalization() 함수를 사용하여 대체하였습니다. 이 결과값을 디코더의 멀티 헤드 어텐션의 입력값으로 대입합니다. 인코더와 마찬가지로 멀티헤드어텐션, FFNN을 거쳐 정규화한뒤 나온 벡터를 linear와 softmax를 거쳐 문장에서 위치할 곳을 확률에 따라 택해주게 됩니다. 이 과정을 반복하게 되면 요약문의 결과값이 나오게 됩니다.

(6페이지)

Gui에 대한 설명문.

(7페이지)

저희는 신조어, 비속어, 등으로 초기 학습 데이터의 퀄리티가 낮아지는 것을 방지하기 위해 최대한 정제되고 정결한 데이터를 찾게 되었고 약 2,000개의 BBC 뉴스 Summary 데이터로 초기학습을 진행하기로 결정했습니다. 다만 2225개의 본문과 요약문 쌍이었기에 학습하기 위한 데이터 양으로는 턱없이 부족했고 이마저도 본문과 요약문의 길이 차이가 크지 않았던 점이 저희 인공지능 성능에 큰 영향을 끼쳤을 것이라고 결론 지었습니다. 또한, 본문별로 텍스트 길이가 많이 달랐던 것 역시 프로젝트를 진행함에 있어 걸림돌이 되었습니다.

ASIC 가속기 같은 전문 장비가 아닌 일반 노트북으로 학습하였기에 학습 소모 시간이(대략 1시간 이상)으로 길었고 인공지능 특성상 내부 계산 과정에서 일어나는 일들을 정확히 알 수 없었기에 코드를 수정하는 과정 역시 어려움이 있었습니다.

원문과 비슷한 요약문이 나오지만 기대했던 것만큼 훌륭한 수준은 아닌 원인을 더 짧고 알맞은 데이터, 이미 학습된 모델을 사용했다면 정확도가 향상되었을 것이라는 결론을 도출하였습니다.

이번 프로젝트를 통해 저희 한옥영 팬클럽팀은 딥러닝 분야의 패러다임 변화를 준 트랜스포머의 아키텍처 구조에 대해 공부하고 구현해볼 수 있는 소중한 경험이 되었던 것 같습니다. 감사합니다.