attention (1,2) 기반 재욱리 강의

숙제

- 파인튜닝기법(LoRA 혹은 파라미터 값 의미조사)
- 강사님 RAG 정리
- 2-3페이지 분량으로 정리

참고

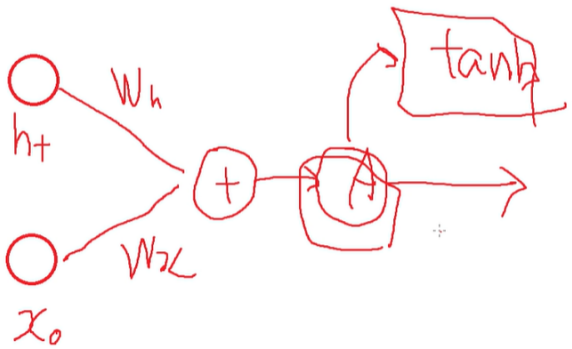
- 경사하강법 - 역전파시 미분의 한 방법

RNN, seq2seq, 어텐션, 트랜스포머, 챗지피티

RNN, 활성화 함수, 기울기 소실

딥러닝 기존 모델 == 이미지 한 컷(story 반영은 불가)

- 그래서 나온 RNN(순환신경망)
  - **rnn은 코난 정주행이다!**
  - rnn 구조로 인풋이 두개(이전 기억*가중치 + 현재 입력 데이터*가중치)



- Activation Func
  - 시그모이드값 0~1 사이
  - 탄젠트 함수 -1 ~1 사이
    - 주로 RNN에서 쓰는 방법
    - 미분 최댓값이 1 임
    - 즉 RNN 층이 깊을수록 예를 들어 0.9를 100번 미분하면?
      - 너무 작은 값이 됨
      - **기울기 소실**
      - 즉, 장기기억 무리

  - ReLU
    - 음수일 땐 0, 양수일 땐 자기자신 값
    - 미분하면 1이 됨
      - 1차식 미분시 1이 나옴
    - **역전파시 미분값 항상 1로 유지됨**

attention

Attention 도입 배경 seq2seq

- seq2seq 구조
  - 인코더, 디코더 구조, N:M 구조
  - RNN 구조임
  - RNN 의 한계
    - 입출력 길이가 N:N이어야 함
      - 예를 들어 한국어와 영어 번역은? n:n이 아님
    - 만화의 경우 최근화만 기억(전반 context 이해 못함)

- **seq2seq는 context vector(요약본 계속 업데이트된 최종 요약본)을 토대로 예측을 하는 모델임**
  - **context vector의 역할은 장기기억 문제를 어느정도 해결하기 위함!**
  - 즉 단어 갯수를 안맞춰도됨
  - 그러나 근본이 RNN 구조이기에 장기기억에 취약함
  - **요약본이란? 전체 내용을 다 못보는 걸 의미 == 성능저하**
  - **그리고 모든 입력 단어의 중요도를 동등하게 봄**
    - 즉 코난의 모든 에피소드를 모두 같은 중요도로 보면 중요한 에피소드나 떡밥을 기억을 못한다! 🎮

Attention 원리

- **핵심은 "코사인 내적으로 중요한 것만 본다!"**
- 코사인 내적이란?
  - 두 벡터를 곱한 길이 * 코사인
  - 같은 방향을 가르키면 1이 됨
  - 우린 두 벡터를 곱한 길이는 구할 수 있음
  - **즉, 코난 각 화(요약본)마다 내적(1에 가까운 애들만 가중치를 줘서 공부한다)**
  - **즉, 디코더가 출력 단어를 예측할 때 입력 시퀀스의 특정 부분에 집중한ㄷ**

Transformer

어텐션 기반의 트랜스포머

- **트랜스포머는 multi head attention이다!**
  - **즉 N개의 화(만화 에피소드)가 있으면 N**2개를 본다 !**
  - 순서, 전체 내용 등 다 아는 어마어마한 놈 🙂
  - 여러 관계를 다 따짐

    - gpu만 그래서 커버 가능
- seq2seq는 대학원생 한명이 순차적으로 데이터를 다 읽는 방식이면 트랜스포머는 중학생 한 30명 모아서 넌 1화 읽어 넌 2화 읽어 이렇게 일을 하는 것
    - 동시 계산!! 그래서 GPU 필요
- 만화책 비유
    - query란 12화를 예상합니다란 상황
    - key란 1,2,3화
    - value: 1화의 값(내용), 2화의 내용
    - **그래서 중요한 key값만 걸러내서 value가져와 decoder로 답변 생성**

Chat GPT

gpt == generated pre-trained transformer

- **챗 지피티란 DECODER ONLY 모델이다!**
    - 그래픽 카드를 엄청 준비해놔 대기업에서
    - 그리고 인코더에 엄~청나게 많은 DATA를 학습 시켜둬
    - 그리고 프롬프터로 현재 상황을 지정해(query))
    - 그리고 단어확률을 계산해
    - 즉! decoder를 우리가 쓰는 것

masked self attention

- **내적할 때 만화 1화~10화까지만 보고(11화는 mask하고) 11화를 예측한다**
- softmax(확률 분포로 만드는 함수)로 예측
- droput 구조 있음(다리 하나씩 끊기)
- mask만 빼면 self attention