

딥러닝을 딥하지 않게 사용해서 자연어 검색 만들기

Langcon 2023 고석현 - Toss

저는

Pikicast
세상을 즐겁게

- Data Lab

이미지, 검색, 추천



Company.AI - co-founder

자연어 대화엔진

Naver에 M&A로 exit

Naver **Clova** - AI Software Platform

Hyper Clova - GPT3 inference optimization

Naver Cloud - 대화엔진 SaaS



toss - ML Platform Team

금융 AI

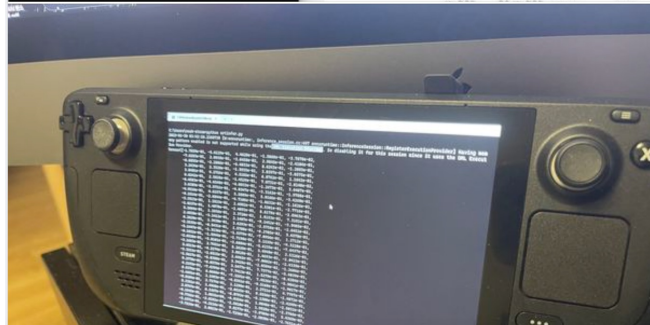
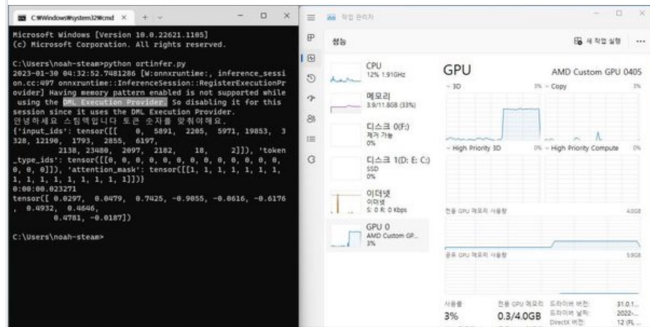


고석현

1월 30일 오전 5:33 · 🌐

최근 예약한 스티픽을 배송 받았습니다. 그래서 모델을 돌려봤습니다.
스티픽은 약 2테라플롭스쯤 되는 GPU 를 가진 커스텀 AMD APU 입니다.

ROCm 을 생각하다가 directx 12 기반 MS Direct Machine Learning을 활용했습니다. 대상 모델은 bert base급 임베딩 모델이고 약 16 ~ 128 개의 토큰을 가지고 추론했습니다. ... 더 보기



관심사는

챗봇 1만 개의 모델 서빙하기: AI 서비스 어디까지 해봤니

Hyper clova 거대 모델 기반의 ML 서빙과 성능 최적화 (GPT-3)

케어콜 - 코로나로부터 우리를 지키는 믿음직한 AiCall

물 흐르듯 자연스러운 ML 서비스 만들기 - MLOps 를 넘어 서비스로

성능 최적화와 **돈이 되는** 머신 러닝에 관심

딥러닝은 무엇일까요

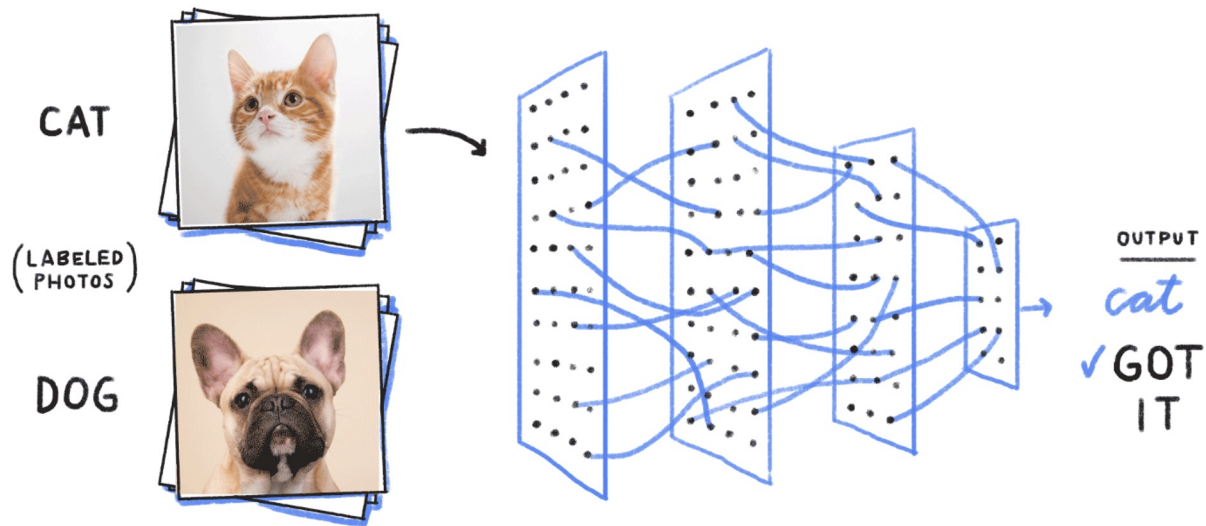
복수의 층으로 구성된 NN를 사용하는 모델과 해당하는 방식

딥러닝은 무엇일까요

비선형 데이터의 집합에 선형성을 부여해주는 projection matrix의 학습 방식

딥러닝은 무엇일까요

feature의 정의를 스스로 하는 모델 ?



딥러닝은 무엇일까요

지금은 Open AI로 이동한 전 Tesla Head of AI 의 Andrej Karpathy 표현을 빌리면
새로운 개발 패러다임에 적합한 도구-> 양질의 데이터는 코드를 대체 하는 수단



Andrej Karpathy

Nov 12, 2017 · 9 min read · [Listen](#)



Software 2.0

I sometimes see people refer to neural networks as just “another tool in your machine learning toolbox”. They have some pros and cons, they work here or there, and sometimes you can use them to win Kaggle competitions. Unfortunately, this interpretation completely misses the forest for the trees. Neural networks are not just another classifier, they represent the beginning of a fundamental shift in how we develop software. They are Software 2.0.

그럼 검색은 무엇일까요

(IR) Information Retrieval ?

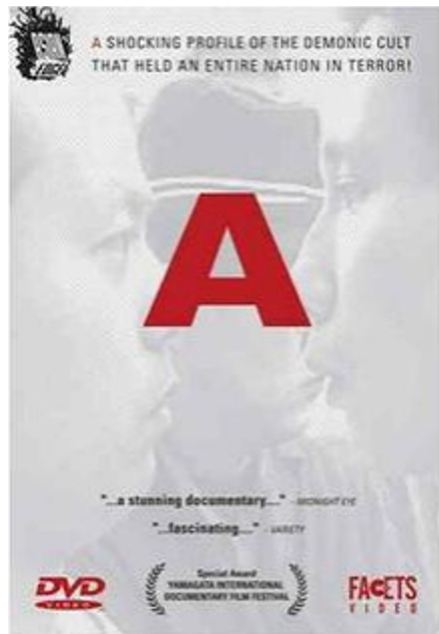
(내 의도와) 적합한 정보를 골라내는것

KR

a 라는 영화 알아?



제가 알아듣기 어려운 질문 같습니다. "a"라는 제목의 영화가 무엇인지 더 자세한 정보를 제공해 주시면 대답해 드릴 수 있을 것 같습니다.



그럼 검색은 무엇일까요

정보가 (지나치게 많아서) 오염되거나 유저에게 적합한 결과가 아니라면
좋은 검색이 아닌 것. 검색의 본질은 결국 결국 추천

KR

You know the movie named A?



I'm sorry, but I would need more information to provide an accurate answer. There are many movies with titles that start with the letter A, and without additional context or details, I'm not sure which movie you are referring to. Could you please provide more information or clarify your question?



KR

movie name is "A"



I'm still unsure which movie you are referring to, as there are many movies with a one-word title "A." Could you provide any additional information about the movie such as the director, main actors, or a brief plot summary? This would help me to better understand which movie you are asking about.



검색은 결국 유저의 의도를 자연어로 입력 받는 추천

검색을 잘하는 방법은 정말 많지만

조금만 열심히 하고 잘하고 싶습니다.

10:26



취소

검색어 입력

신라면만드는회사



"신라면만드는회사" 종목 검색하기

토스증권

"농심" 종목 검색하기

토스증권

"농심홀딩스" 종목 검색하기

토스증권

"롯데칠성" 종목 검색하기

토스증권

그래서 쉬운 딥러닝을 써서 조금만 열심히 하고 잘하고 싶습니다.

소프트웨어 2.0

나는 때때로 사람들이 뉴럴넷을 "기계학습 도구 상자의 또 다른 도구"로 언급하는 것을 본다. 뉴럴넷은 장단점이 있다. 여기 저기에서 동작하고, 때로는 캐글 경쟁에서 승리하는데 사용할 수도 있다. 그러나 이 해석은 완전히 나무만 보고 숲을 보지 못하는 것이다. 뉴럴넷은 단지 또 다른 종류의 분류기가 아니다. 그것은 우리가 소프트웨어를 작성하는데 있어 근본적인 변화의 시작을 나타낸다. 그것은 **소프트웨어 2.0**이다.

쉽게 이용할 수 있다.

나는 '딥러닝은 알파하드'라는 농담을 좋아한다. 이것은 당신이 유용한 무언가를 얻기 위해 박사 학위가 필요하지 않다는 것이다. 기본 개념은 기초 선형 대수, 미적분, 파이썬 및 CS231n의 일부 강의 정도이다. 물론 오랜 경험에서 얻을 수 있는 전문 지식과 직관력의 양은 아주 많다. 보다 정확한 말은 소프트웨어 2.0 스택은 쉽게 이용할 수 있지만, 쉽게 마스터 할 수는 없다는 것이리라.

그것은 당신보다 낫다.

마지막으로, 그리고 가장 중요하게, 뉴럴 네트워크는 나나 당신이 현재 가치있는 업종의 상당 부분에서 만들 수 있는 것보다 더 훌륭한 코드라는 것이다. 그것은 현재의 이미지 / 비디오, 음성 / 담화와 별로 관련이 없다.

조금만 열심히 하기 위해서

논문을 찾아봤습니다.

Dense Passage Retrieval for Open-Domain Question Answering

EMNLP 2020, facebook research

- 일반적인 Open Domain QA의 경우 BM25와 같은 희소 벡터 모델을 통해 문서 후보를 선택하고 이후 밀집 벡터 검색으로 passage 단위의 후보를 찾음
- 하지만 우리는 밀집 벡터 모델만으로 검색이 구현가능 하다는 것을 보임
- 단순하게 구성된 bi encoder 구조에서 튜닝된 Lucene-BM25 보다 9~19% 우수한 SOTA의 성능을 보였다.
- 충분히 적은 학습 데이터로도 잘 동작 하였다.

토스엔 검색팀이 없다...

검색 하고 싶은 사람 손 ! 해서 슬랙에서 모인 여러팀의 사람들이 자율적으로 개발
최소 공수, 최단 주기, 데이터 없음, 기존 로그 당연히 없음, 검색 대상 정의 없음

하지만 유저의 기대치는 매우 높음

그리고 다들 본업도 따로 있음, 그래서 처음은 조금만 열심히 해야 했습니다.

2 주 안에 서비스 통합 검색을 만들기로 와우!

더 자세한 내용은 Langcon 2023 에서 !