Hugging Face NLP

자연어 처리

Hugging Face

자연어 처리(Natural Language Processing)

자연어

인간의 의사 소통을 위해 자연스럽게 발달한 언어를 의미.

자연어 처리는 컴퓨터가 인간의 언어를 이해하고 해석 및 처리하기 위한 기술을 의미.

순환 신경망(RNN:Recurrent Neural Network)과 장단기 기억(LSTM:Long Short Term Memory) 모델은 시퀀스 데이터를 처리하는데 효과적 이었으나 순차적 처리 방식으로 인해 병렬 연산이 제한되어 대규모 처리 시 계산 효율이 떨어짐.

트랜스포머의 등장으로 입력 전체를 병렬로 처리해 데이터 처리 속도를 향상시켰다.

이로 인해 대규모 데이터를 활용한 효율적인 사전학습이 가능해졌고, 언어의 문법과 어휘를 깊이 이해하는 고성능 언어 모델을 구축할 수 있게 되었다.

텍스트 분류 : BERT 모델을 사용하여 텍스트를 여러 카테고리로 분류.

요약문 생성 : BART 모델을 활용하여 긴 문서의 핵심 내용을 추출하고 간결한 요약문을 생성.

질의 응답 : RoBERTa 모델을 이용해 주어진 문서에서 질문에 대한 정확한 답변을 추출.

기계 번역 : T5 모델을 사용하여 한 언어에서 다른 언어로 텍스트를 번역하는 방법 및 생성.

텍스트 생성 : 최신 대규모 모델인 LLaMA-3.1을 활용하여 텍스트 생성.

wandb(Weights andBias) 설치

1. 회원 가입

: https://wandb.ai/home

2. API

https://app.wandb.ai/authorize

To find your API key for Weights and Biases (W&B):

Log in to W&B at https://wandb.ai/authorize.

Alternatively, access your profile:

Click your user profile in the upper right corner.

Select “User Settings.”

Scroll to the “Danger Zone” section.

Click “Reveal” next to “API Keys.”

텍스트 분류: BERT

텍스트 분류(Text Classification)는 입력 텍스트를 미리 정의된 범주나 레이블로 할당하는 방법이다.

예를 들어 Email 시스템에서는 텍스트 분류 기술을 활용해 스팸과 정상 메일로 구분한다.

또한 뉴스 기사나 문서를 주제별로 자동 분류 하거나 고객 문의 메일을 유형별로 분류하는 등 다양한 방면에서 텍스트 분류 기술을 활용할 수 있다.

텍스트 분류 모델의 성능은 데이터의 질과 양, 전처리 방법, 모델 아키텍쳐 등 다양한 요인에 의해 결정 된다.

따라서 실제 응용 분야에 맞는 적절한 전략을 수립하는 것이 중요하다.

또한 텍스트 분류 모델은 편향성과 공정성 문제에 주의해야 하며 모델의 예측 결과를 적절히 해석하고 활용할 수 있어야 한다.

BERT(Bidirectional Encoder Representations from Transformers)는 2018년 구글에서 개발한 대표적인 사전 학습 언어 모델이다.

트랜스포머 아키텍쳐를 기반으로 대량의 말뭉치에서 사전 학습된 BERT는 자연어 처리 분야의 중요한 분기점이다.

요약문 생성: BART

요약문 생성(Summary generation)은 방대한 양의 텍스트 정보를 간결하고 명료하게 압축해 전달하는 부분이다.

텍스트 요약의 종류

1. 추상적 요약(Abstractive summarization)

: 원문 텍스트의 의미를 완전히 이해하고 새로운 문장을 생성해 요약하는 방식.

뉴스 요약, 과학 논문 요약 등 높은 수준의 자연어 이해와 생성 능력이 요구되는 분야에 적합.

2. 추출적 요약(Extractive summarization)

: 원문에서 가장 중요하고 관련성이 높은 문장들을 그대로 추출하는 방식.

특허 문서, 법률 문서, 회의록 등 길고 반복적인 텍스트에서 중요 문장을 추출하는데 유용.

BART(Bidirectional and Auto-Regressive Transformers)

메타(facebook)에서 개발한 언어 모델로 인코더-디코더 아키텍쳐를 갖춘 Sequence-to-Sequence 모델로

기존의 순차적인 언어모델과 달리 BART는 양방향 컨텍스트를 활용해 언어 이해 및 생성 능력이 뛰어나다.

질의 응답: RoBERTa

질의 응답(Question answering)은 주어진 지식이나 맥락을 바탕으로 사용자가 제시한 질문에 적절한 답을 제공하는 과제.

질의 응답 종류

1. 추출 질의 응답(Extractive question answering)

: 주어진 지문 내에서 답변이 되는 연속된 문자열 추출.

지문 내에 정답이 존재 하므로 비교적 간단하고 정확한 답변을 제공할 수 있으나 지문에 명시적으로 드러나지 않은 정보에 대해서는 답변이 어려움.

2. 생성 질의 응답(Generative question answering)

: 질문과 지문을 입력받아 새로운 답변을 생성

지문외의 외부 지식을 활용해 답변할 수 있지만 생성된 답변의 정확성과 일관성을 보장 하기가 쉽지 않다.

RoBERTa(A Robustly Optimized BERT Pretraining Approach)

BERT모델의 성능을 개선하기 위해 메타의 FAIR에서 제안한 변형 모델

기계 번역: T5

기계 번역(Machine translation)은 컴퓨터 프로그램을 활용해 한 언어의 텍스트를 다른 언어의 텍스트로 자동 변환하는 기술.

기계 번역 종류

1. 통계적 기계 번역(Statistical Machine Translation : SMT)

원문과 번역문 쌍을 기반으로 단어 순서와 언어 패턴을 인식해 학습한다.

단어 시퀀스를 불연속적인 기호로 취급. 대용량의 데이터가 필요하다.

2. 신경망 기계 번역(Neural Machine Translation: NMT)

심층 신경망 모델을 사용해 번역문과 단어 시퀀스 간의 관계를 학습한다.

단어를 밀도 벡터로 표현해 연속적인 공간에 매핑 한다. 문맥을 고려한 번역이 가능해 문장 전체의 일관성과 유창성이 크게 향상 된다.

T5(Text-To-Text Transfer Transformers)

구글에서 개발한 언어 모델로 인코더와 디코더로 이루어진 Sequence-To-Sequence 모델로 자연어 처리를

Text-To-Text형태의 데이터로 변환하고 이를 Sequence-To-Sequence 문제로 인식해 해결.

텍스트 생성: LLaMA-3.1

텍스트 생성(Text Generation)은 주어진 입력 텍스트를 기반으로 새로운 텍스트를 만들어 내는 기술이다.

단순히 주어진 입력을 바탕으로 텍스트를 제작하는 것을 넘어서 맥락을 이해하고 적절한 응답을 생성하는 복잡한 과정을 포함한다.

허깅페이스 트랜스포머 라이브러리에는 크게 Sequence-to-sequence 기반 모델과 인과적 언어 모델(Casual language model)을 제공한다.

Sequence-to-Sequence Model

- 대표적인 모델 : Transformer, MASS, BART, T5
- 구조 : 인코더와 디코더로 구성
- 특징 : 양방향 모델로 입력 텍스트 전체를 고려해 출력을 생성
- 주요 응용분야 : 기계 번역, 텍스트 요약, 질의 응답 시스템

인과적 언어 모델

- 대표적인 모델 : GPT, LLaMa, ChatGPT, Gemma, PaLM
- 구조 : 단일 디코더로 구성
- 특징 : 단방향 모델로 이전 토큰들을 기반으로 다름 토큰을 예측
- 주요 응용 분야 : 대화 시스템, 텍스트 생성, 다음 단어 예측

텍스트 생성: LLaMA-3.1

LLaMA(Large Language Model Meta AI)는 메타에서 개발 및 공개한 대규모 언어 모델 시리즈로 LLaMA-3는 2024년 4월에 공개.

현재 8B(80억)와 70B(700억) 개의 매개변수를 갖는다.

LLaMA-3는 15조개 이상의 토큰으로 학습으로 30개 이상의 언어로 된 고품질 데이터가 포함되어 있어, 모델의 다국어 처리 능력이 크게 향상.

LLaMA-3의 학습 데이터에는 일반 지식에 해당하는 토큰이 약 50%, 수학 및 추론 토큰이 25%, 코드 토큰이 17%, 다국어 토큰이 8%.

LLaMA-3 모델은 규모와 성능으로 인해 상당한 컴퓨팅 자원을 요구한다.

가장 작은 모델 조차도 80억개의 매개변수를 보유하고 있으며 이를 운용하기 위해서는 최소 20GB의 GPU Memory가 필요하다.

더욱이 학습 과정에서는 기울기 계산이 추가적인 메모리가 소요되어 실제 학습은 매우 제한적이다.

따라서, 모델을 직접 학습시키는 대신 지시학습(Instruction tunnung)된 LLaMA-3 모델을 불러와 텍스를 생성하는 방법으로 실습한다.

텍스트 생성: LLaMA-3.1

LLaMA3 의 사용 권한 및 Token 받기

1. **Hugging Face 계정 로그인 및 라이선스 동의**
   - Hugging Face에 접속하여 계정으로 로그인합니다.
   - LLaMA 3 모델 페이지로 이동합니다. 예: meta-llama/Meta-Llama-3.1-8B-Instruct

     https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
   - 모델 페이지에서 **“Access request”** 또는 **“Request Access”** 링크(첫 화면의 2번째 파트)를 찾아 클릭 합니다.
   - 제시되는 라이선스 약관을 읽고 동의합니다.
2. **접근 승인 대기**
   - 라이선스에 동의한 후, Meta의 승인을 기다려야 합니다(약 15분 소요).

텍스트 생성: LLaMA-3.1

Access Token 생성

1.허깅페이스 로그인 후, 우측 상단의 프로필 사진을 클릭합니다.

2.드롭다운 메뉴에서 **Settings**를 선택합니다.

3.설정 페이지에서 좌측 메뉴의 **Access Tokens** 항목을 클릭합니다.

4.**New Token** 버튼을 클릭하여 새 토큰을 생성합니다.

- **Token Name**: 토큰 이름을 지정합니다(예: LLaMA3-learning).
- **Gated Repos Status**가 Accept인지 확인
- **Fine Grainded** 선택
- **Search Repos**를 선택하여 meta-llama를 선택