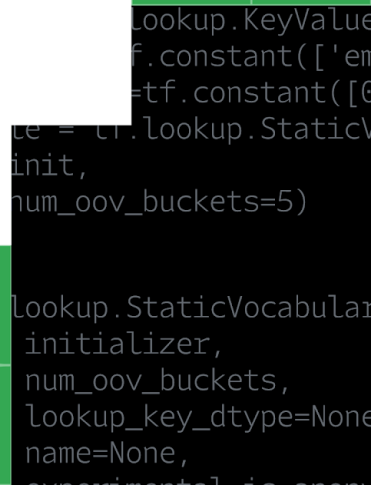
Google Developer Student Clubs

3주차 2조

팀원: 황동욱, 권도혁, 송성근



(황동욱) 알고리즘 정리

🤔 BERT란

BERT(Bidirectional Encoder Representations from Transformers)는 Google AI에서 2018년에 발표된 자연어 처리(NLP)를 위한 전이 학습 모델입니다. BERT는 그 이전의 여러 자연어 처리 모델들과는 달리 양방향으로 문맥을 고려하기 때문에 높은 성능을 보이며, 여러 NLP 태스크에서 최첨단 성능을 기록했습니다.

BERT의 주요 특징 및 아이디어는 다음과 같습니다:

1. **양방향 Transformer**: BERT는 Transformer의 인코더 구조를 사용하여 입력 텍스트의 양방향 문맥을 학습합니다. 이를 통해 각 단어가 주어진 문장에서의 전체 문맥을 잘 이해할 수 있게 됩니다.
2. **전이 학습**: BERT는 큰 코퍼스(예: Wikipedia)에서 미리 훈련된 모델을 사용하여 특정 NLP 태스크(예: 질문 응답, 감정 분석)를 위해 더 적은 데이터와 함께 미세 조정(fine-tuning)할 수 있습니다.
3. **Masked Language Model(MLM) 학습**: BERT의 훈련 중 한 가지 주요 기법은 MLM입니다. 이때 일부 단어는 마스크(mask) 처리되며, 모델은 이 마스크된 단어를 예측하려고 합니다. 이를 통해 모델은 문장 내의 모든 단어에 대해 깊은 문맥적 이해를 얻게 됩니다.
4. **다양한 크기의 버전**: BERT는 여러 가지 크기의 버전이 있으며, 예를 들면 BERT-Base, BERT-Large 등이 있습니다. 큰 버전은 더 많은 모델 파라미터를 가지며, 더 정확하지만 연산 비용이 큽니다.
5. **다양한 언어 지원**: BERT는 다양한 언어를 지원하며, 여러 언어에 대한 학습 데이터로 사전 훈련된 다양한 버전의 모델이 공개되어 있습니다.

요약하면, BERT는 Transformer 아키텍처를 기반으로 한 양방향 문맥 인코딩 능력과 전이 학습 능력을 결합하여 여러 NLP 태스크에서 뛰어난 성능을 제공하는 모델입니다.

🤔 전이 학습이란

전이학습(Transfer Learning)은 이미 한 분야나 태스크에서 학습된 모델의 지식을 다른, 관련된 분야나 태스크에 적용하여 학습 효율을 향상시키는 기법을 말합니다. 즉, 하나의 태스크에서 얻은 지식을 다른 태스크에 '전이'하는 것입니다.

전이학습의 주요 특징 및 이점은 다음과 같습니다:

1. **데이터 부족**: 적은 양의 데이터로도 좋은 성능의 모델을 학습시킬 수 있습니다. 특히, 특정 분야나 태스크에 대한 충분한 학습 데이터가 부족한 경우, 전이학습은 큰 도움이 됩니다.
2. **학습 속도 향상**: 이미 학습된 모델의 가중치와 구조를 기반으로 새로운 태스크를 학습하기 때문에, 학습 속도가 상대적으로 빠를 수 있습니다.
3. **효과적인 학습**: 복잡한 모델, 특히 딥 러닝 모델의 경우, 처음부터 학습하는 것보다 전이학습을 통해 초기 가중치를 설정하는 것이 종종 더 효과적일 수 있습니다.

전이학습의 대표적인 예로, 이미지 분류를 위해 사전에 대규모 데이터셋(예: ImageNet)에서 학습된 심층 신경망 모델을 사용하여, 다른 이미지 분류 태스크나 객체 탐지 태스크 등에 적용하는 경우를 들 수 있습니다.

자연어 처리 분야에서도, BERT, GPT 등의 큰 모델이 특정 데이터셋에서 미리 학습된 후, 다양한 NLP 태스크에 미세조정(Fine-tuning)을 통해 전이학습되는 방식이 널리 사용되고 있습니다.

🤔Fine-tuning 이란

Fine-tuning(미세조정)은 전이학습의 한 형태로, 이미 학습된 모델(pre-trained model)의 파라미터를 특정 태스크에 맞게 추가적으로 학습시키는 과정을 말합니다. 미세조정은 기본적으로 큰 데이터셋에서 사전 학습된 모델의 지식을 활용하여, 작은 데이터셋으로도 효율적인 학습을 가능하게 하며, 일반적으로 학습 속도와 성능을 향상시킵니다.

예시:

1. **이미지 분류**:
   - **사전 학습**: ImageNet이라는 대규모 이미지 데이터셋에서 학습된 신경망 모델이 있다고 가정합시다. 이 모델은 천 개의 다양한 카테고리를 분류하는 데에 사용되었습니다.
   - **미세조정**: 이제 우리는 이 모델을 활용하여 강아지 품종을 분류하는 작은 데이터셋으로 학습시키고자 합니다. 이 때, 모델의 마지막 계층을 특정 태스크에 맞게 수정하고, 전체 모델 또는 일부 계층을 추가적으로 학습시킵니다. 이러한 과정을 통해 강아지 품종을 분류하는 모델을 효과적으로 학습할 수 있습니다.
2. **자연어 처리**:
   - **사전 학습**: BERT는 대규모 텍스트 데이터셋(예: Wikipedia)에서 학습된 모델로, 다양한 언어 모델링 태스크에 사용되었습니다.
   - **미세조정**: 이제 이 BERT 모델을 특정 NLP 태스크, 예를 들어 감정 분석에 맞게 학습시키고자 합니다. BERT 모델의 출력 계층을 수정하고, 추가적인 학습을 통해 해당 태스크에 최적화된 모델을 얻을 수 있습니다.

Fine-tuning의 핵심은, 사전 학습된 모델이 가진 일반적인 지식을 활용하면서, 특정 태스크에 필요한 세부적인 지식을 추가적으로 학습시키는 것입니다.

🤔TF-IDF:

- TF-IDF는 Term Frequency-Inverse Document Frequency의 약자로, 정보 검색과 텍스트 마이닝에서 문서의 중요도를 평가하는 데 사용되는 통계적 방법입니다.
- TF(Term Frequency)는 특정 단어가 문서 내에서 얼마나 자주 등장하는지를 나타내며, IDF(Inverse Document Frequency)는 특정 단어가 얼마나 많은 문서에 등장하는지의 역수를 나타냅니다. TF-IDF는 이 두 값을 곱한 것으로, 문서 내의 단어 중요도를 평가할 때 주로 사용됩니다.

(권도혁) 알고리즘 정리

BERT란 무엇인가?

BERT?

구글에서 개발한 NLP(자연어처리) 사전 훈련 기술이며, 특정 분야에 국한된 기술이 아니라 **모든 자연어 처리 분야에서 좋은 성능을 내는 범용 Language Model**입니다. 11개 이상의 자연어처리 과제에서 BERT가 최첨단 성능을 발휘한다고 하지만 그 이유는 잘 알려져 있지 않다고 합니다. 하지만 BERT는 지금까지 자연어처리에 활용하였던 앙상블 모델보다 더 좋은 성능을 내고 있어서 많은 관심을 받고 있는 언어모델 입니다.

[출처] https://ebbnflow.tistory.com/151

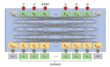
BERT는 학습모델이 아니다!

사전 훈련 (Pre-trained)된 언어모델로 Word2Vec 등을 대신해서 Embedding에 이용

BERT public 코드 제출 결과

Starter Notebook: Ranked Predictions with BERT - Version 1 0.515

Succeeded · 15h ago · Notebook Starter Notebook: Ranked Predictions with BERT | Version 1



Huggingface BERT

BERT models directly retrieved and updated from: https://huggingface...

Last Updated: 3 days ago (Version 144)

About this Dataset

This dataset contains many popular BERT weights retrieved directly on Hugging Face's model repository, and hosted on Kaggle. It will be automatically updated every month to ensure that the latest version is available to the user. By making it a dataset, it is significantly faster to load the weights since you can directly attach a Kaggle dataset to the notebook rather than downloading the data every time. See the speed comparison notebook.

The banner was adapted from figures by Jimmy Lin (tweet; slide) released under CC BY 4.0. BERT has an Apache 2.0 license according to the model repository.

Tokenizer

Tokenizer

Huggingface에서 제공하는 대부분의 transformer 모델들은 텍스트를 바로 입력으로 받을 수 없다. 예를 들어 BERT 모델의 경우 문자열(string)로 된 텍스트 그 자체가 아닌, 텍스트를 tokenize한 후 각 token들을 고유한 정수로 바꾼 것, 그러니까 token id들의 리스트(sequence of token ids)를 입력으로 받는다. 게다가 `attention_mask`, `token_type_ids`와 같은 추가적인 데이터를 요구한다.

이때 Huggingface에서는 각 모델별로 tokenizer라는 것을 제공한다. tokenizer는 자신과 짝이 맞는 모델이 어떤 형태의 입력을 요구하는지를 알고 있어, tokenizer를 이용하면 텍스트를 전처리, 가공해 모델의 입력값(model input)을 만들 수 있다. 우린 그냥 tokenizer의 출력을 그대로 모델에 넣어주기만 하면 된다.

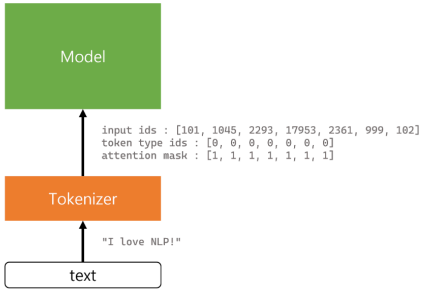


Fig.01 Huggingface Tokenizer

tokenizer는 모델이 받아들일 수 있는 형태로 입력 텍스트를 가공해 준다.

그래서 Huggingface 모델 사용의 제 1단계는 "해당 모델의 tokenizer 불러오기" 이다. 다음과 같이 `AutoTokenizer.from_pretrained()` 메소드에 모델명을 입력하면 모델과 매치되는 tokenizer를 불러올 수 있다. 이를 이용해 `bert-base-uncased` 모델의 tokenizer를 불러오자.

`bert-base-uncased` tokenizer 출력값에 있는 각 항목의 의미는 다음과 같다.

- `input_ids` : token들의 id 리스트(sequence of token id).
- `token_type_ids` : 각 token이 어떤 문장에 속하는지를 나타내는 리스트. BERT는 한 번에 두 문장(sentence A, sentence B)을 입력으로 받을 수 있는데, `bert-base-uncased` tokenizer는 sentence A에 속하는 token에는 0을, sentence B에 속하는 token에는 1을 부여한다.
- `attention_mask` : attention 연산이 수행되어야 할 token과 무시해야 할 token을 구별하는 정보가 담긴 리스트. `bert-base-uncased` tokenizer는 attention 연산이 수행되어야 할, 일반적인 token에는 1을 부여하고, padding과 같이 attention 연산이 수행될 필요가 없는 token들에는 0을 부여한다.

추석연휴간 BERT Code 이해,
추가로 DeBERT & Longformer 에 대한 개념 이해를 목표로 학습할 예정.

[출처] https://heekangpark.github.io/nlp/huggingface-bert

(송성근) 알고리즘 정리

MultipleChoice

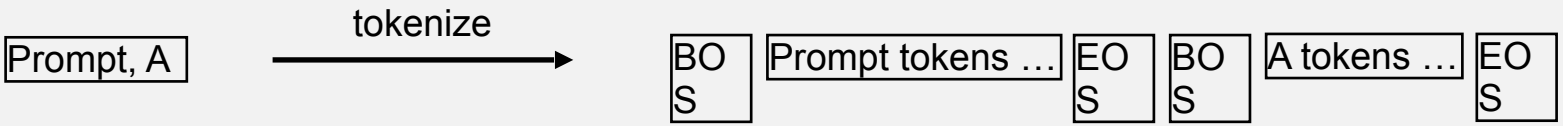
여러 개의 답 찾는 task를 진행하기 위한 모델

기존 두 문장 사이 관계를 찾는 task와 동일하게 각 선택지에 대해서
[문제, 선택지] 형태로 tokenize를 진행시킨다

Label 자체는 하나의 답을 사용하여 학습하고
Prediction시 output에서 결과가 높은 순으로 n개의 답을 선택하는 방식

데이터셋 예시 1
Prompt : What is the term used in astrophysics to describe light-matter interactions resulting in energy shifts in the radiation field?

A. Blueshifting B. Redshifting C. Reddening D. Whitening E.Yellowing



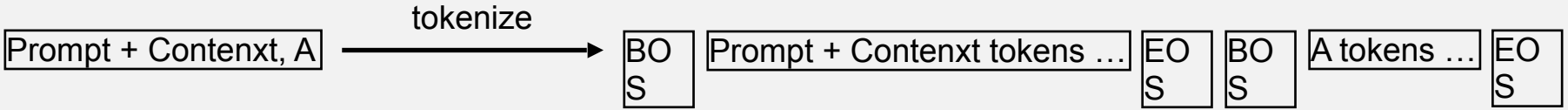
MultipleChoice

데이터셋 예시 2

Prompt : "In relation to Eunice Fay McKenzie's career, which statement accurately reflects her most notable work?“

Context : "Eunice Fay McKenzie (February 19, 1918 – April 16, 2019) was an American actress and singer. She also entertained the troops with her former screen partner, ..”

A. “…”, B. “…”, C. “…”, D. “…”, E. “…”



DeBERTa

기존 BERT 모델에서 2가지 방식으로 성능을 개선

기존 BERT와는 다르게 input으로 받는 token들을 embedding 단에서
content와 position을 각각 인코딩하는 두 개의 벡터로 표현

self attention layer에서 각 word들의 relative position에 따른 가중치 부여 (disentangled attention)

MLM으로 pre-trainin시 masked token을 예측하는 과정에서 absolute position 정보를 추가
(Enhanced Mask Decoder)

A new store opened beside the new mall

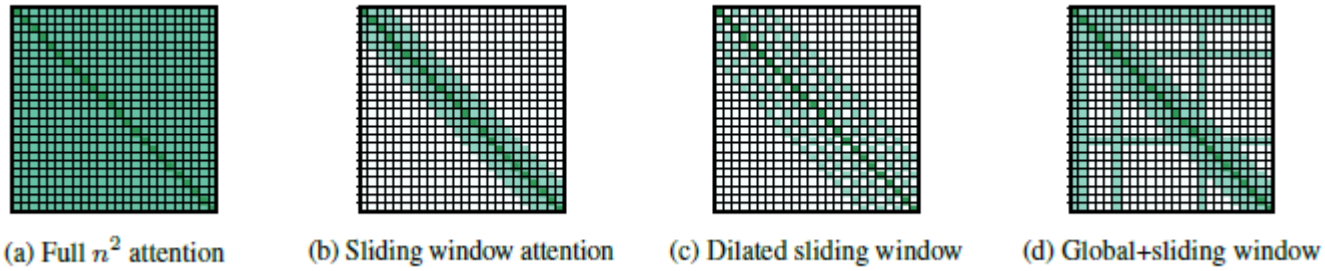
서로 의미가 비슷하지만 전반적인 단어의 위치를 고려하였을 때 다른 장소를 의미한 것을 알 수 있음 (absolute position)

사용 모델

Longformer

Sequence 길이에 선형적으로 비례하는 attention 메카니즘

BERT 모델은 token의 사이즈가 512를 넘어가면 일부 문장에 대해서 손실이 발생
긴 문서에 대한 task를 진행하기 위해 연산량을 효율적으로 변경



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

자신 근처 w개의 토근에 대해서 window 구성 attention 진행 (b)
n개씩 띄엄띄엄 형태로 attention 진행 (c)

→ 애네를 이용해 global attention 진행 (d)