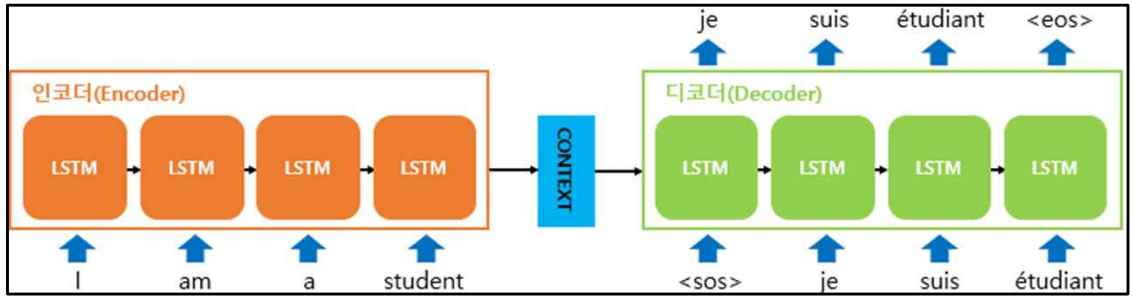
Transformer와 GPT

1. 구글 Transformer 모델

인코더와 디코더 (Seq2Seq) 모델?

인코더-디코더(Encoder-Decoder) 모델은 시퀀스 데이터를 입력 받아 또 다른 시퀀스를 출력하는 다양한 자연어 처리 작업에 사용되는 딥러닝 아키텍처이다. 이 모델은 기계 번역, 텍스트 요약, 질문 응답 시스템 등 시퀀스 간의 변환이 필요한 작업에서 주로 활용된다.

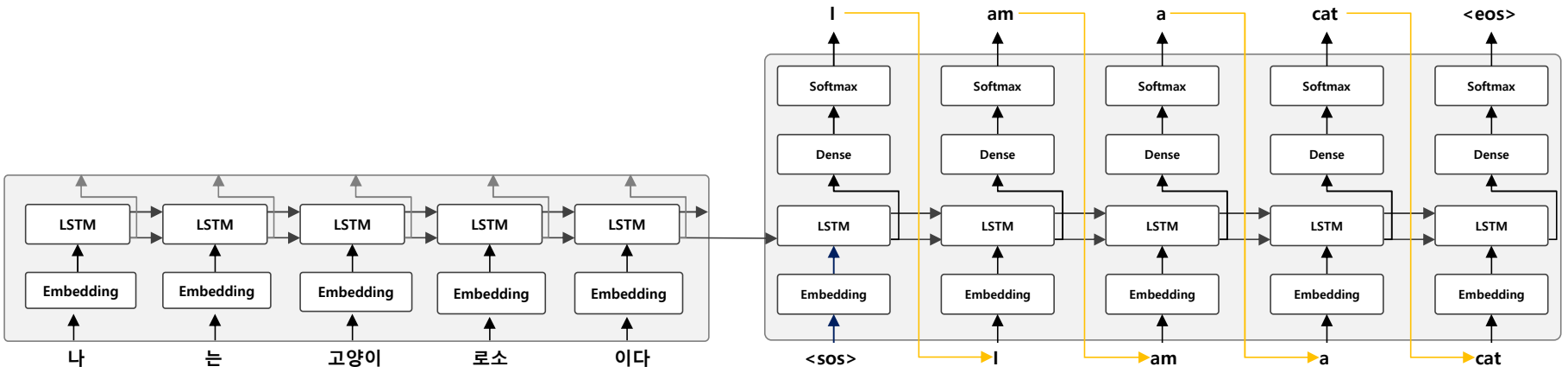


기본 구조

인코더-디코더 모델은 두 개의 주요 구성 요소로 이루어짐

- **인코더(Encoder)**: 입력 시퀀스를 받아 이를 고정된 크기의 벡터 표현으로 변환. 이 벡터는 입력 시퀀스의 중요한 정보를 압축하여 담고 있다.
- **디코더(Decoder)** : 인코더에서 생성된 벡터를 입력으로 받아, 목표 출력 시퀀스를 생성. 이 과정에서 디코더는 이전 단계에서 생성된 출력(예: 번역된 단어들)을 이용하여 다음 단어를 예측한다.

인코더 디코더(seq2seq)의 구성



인코더 디코더

동작 방식

1.입력 처리(인코더 부분):

인코더는 주어진 입력 시퀀스(예: 문장)를 처리하여 각 단어에 대한 벡터 표현을 생성한다.
이 벡터들은 주로 RNN, LSTM, GRU 또는 Transformer와 같은 아키텍처를 사용하여 처리된다.
인코더의 마지막 상태는 입력 시퀀스 전체의 정보를 요약한 하나의 고정된 벡터(컨텍스트 벡터, Context Vector)로 압축된다.

2. 출력 생성(디코더 부분):

- 디코더는 인코더에서 생성된 컨텍스트 벡터를 받아, 출력 시퀀스의 첫 번째 단어를 예측한다.
- 예측된 첫 번째 단어를 기반으로 다음 단어를 예측하며, 이 과정이 반복된다.
- 디코더는 출력 시퀀스 전체가 생성될 때까지 이 과정을 계속한다.

Attention 메커니즘 (앞 장 pdf 파일 265 Page 참조)

시퀀스 데이터(예: 텍스트)에서 중요한 부분에 집중하여 더욱 효과적으로 정보를 처리하는 방법이다 특히 자연어 처리에서 매우 중요한 역할을 하며, 텍스트의 문맥을 더 잘 이해할 수 있도록 돕는다.

1. 순환 신경망의 문제점

기존의 순환 신경망(RNN)이나 LSTM은 시퀀스를 순차적으로 처리한다. 하지만 이 방식은 긴 시퀀스에서 과거의 정보가 희석되거나 잊혀질 수 있다는 문제가 있다. 또한, 이 모델들은 모든 입력을 동일한 비중으로 처리하기 때문에, 특정 단어들이 문맥상 더 중요한 경우에도 이를 구별하지 못하는 한계가 있다.

2. Attention 메커니즘의 아이디어

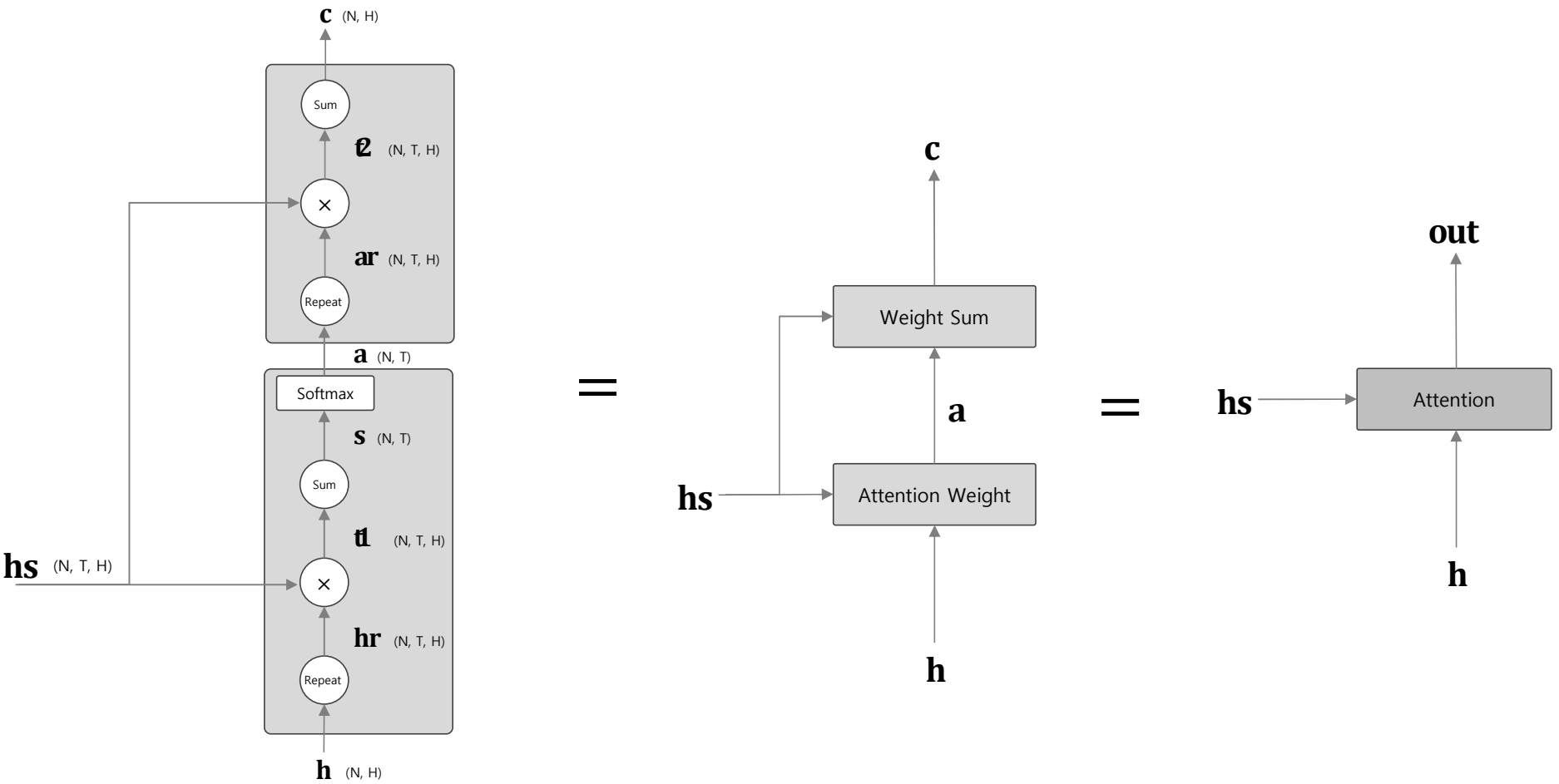
Attention 메커니즘의 기본 아이디어는 모든 입력 단어에 대해 "집중해야 할 단어"를 선택하고, 그 중요도를 동적으로 계산하는 것이다. 이를 통해 모델은 특정 단어들이 다른 단어들에 비해 더 중요한지 여부를 학습한다.

3. Self-Attention

Self-Attention은 입력 시퀀스 내의 각 단어에 대해, 다른 모든 단어 들과의 연관성을 계산하는 방식이다. 이를 통해 문장의 각 단어가 문맥적으로 얼마나 중요한지를 평가한다.

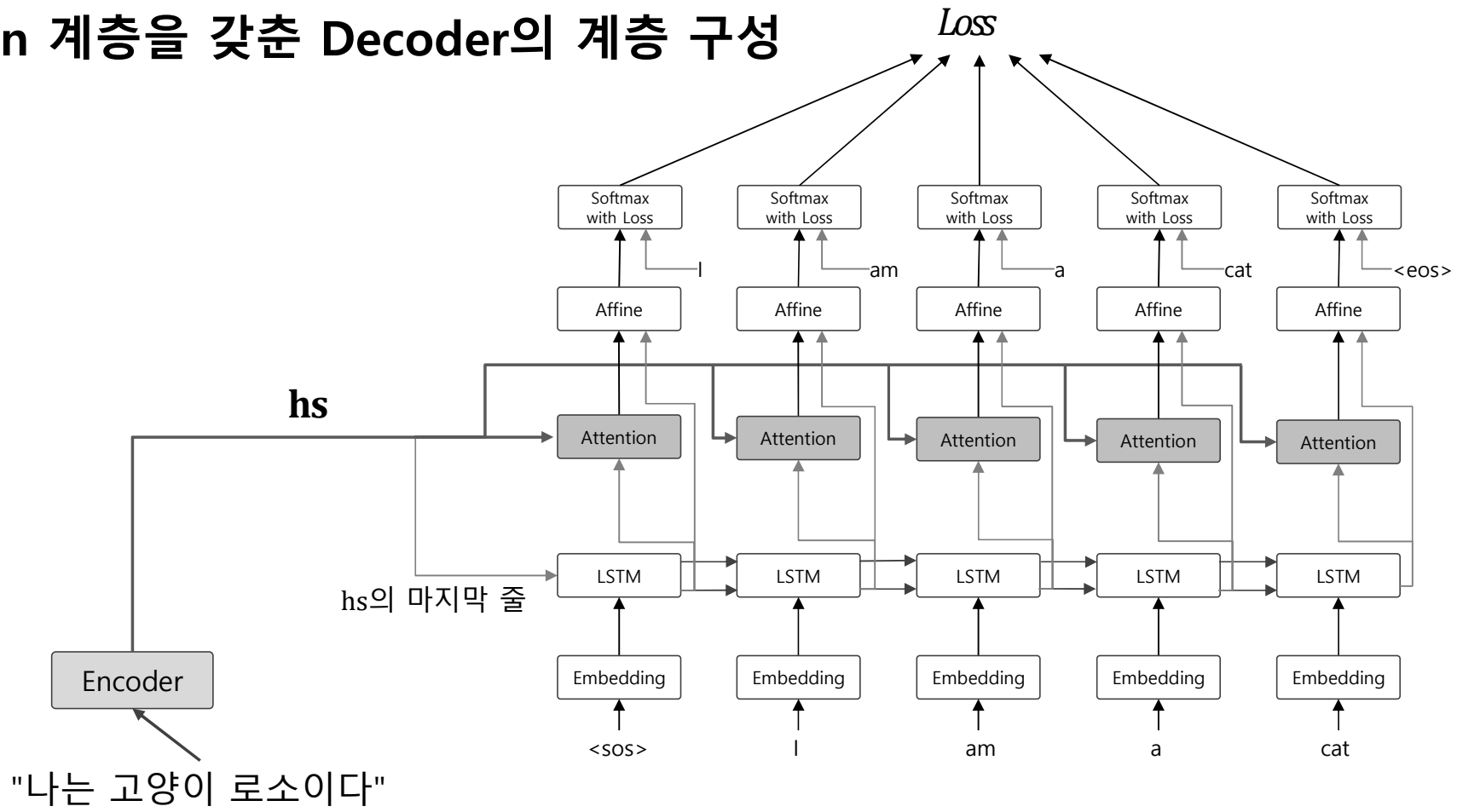
1. 구글 Transformer 모델

맥락 벡터를 계산하는 계산 그래프



1. 구글 Transformer 모델

Attention 계층을 갖춘 Decoder의 계층 구성



1. 구글 Transformer 모델

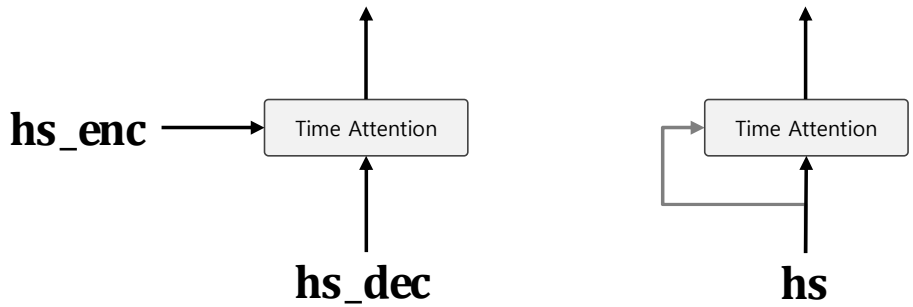
Transformer 모델은

자연어 처리(NLP) 분야에서 혁신적인 성과를 이룬 딥러닝 모델로, 2017년 구글 연구진이 발표한 "Attention is All You Need" 논문에서 처음 소개되었다.

Transformer 모델의 주요 특징과 구성 요소는 다음과 같다

1. **Attention 메커니즘**: Transformer 모델의 핵심은 "Self-Attention" 메커니즘이다. 이 메커니즘은 문장 내의 각 단어가 다른 단어 들과의 관계를 학습하여 문맥을 이해하는 데 도움을 준다. Self-Attention은 각 단어에 대해 문장의 다른 모든 단어 들과의 연관성을 계산하여, 중요한 단어들에 더 많은 비중을 두는 방식으로 동작한다.

왼쪽이 일반적인 어텐션, 오른쪽이 셀프어텐션



1. 구글 Transformer 모델

2. 병렬 처리 : RNN이나 LSTM과 같은 기존 모델들은 순차적으로 데이터를 처리하기 때문에 병렬화가 어렵지만, Transformer 모델은 병렬 처리가 가능하여 학습 속도가 훨씬 빠르다. 이는 특히 대규모 데이터셋을 다룰 때 큰 장점으로 작용한다.

3. 인코더-디코더 구조 : Transformer 모델은 인코더와 디코더라는 두 개의 주요 모듈로 구성된다. 인코더는 입력 문장을 처리하여 의미 있는 표현을 만들어내고, 디코더는 이 표현을 바탕으로 출력 문장을 생성한다. 번역 작업과 같은 시퀀스-투-시퀀스 문제에서 주로 사용된다.

1. 구글 Transformer 모델

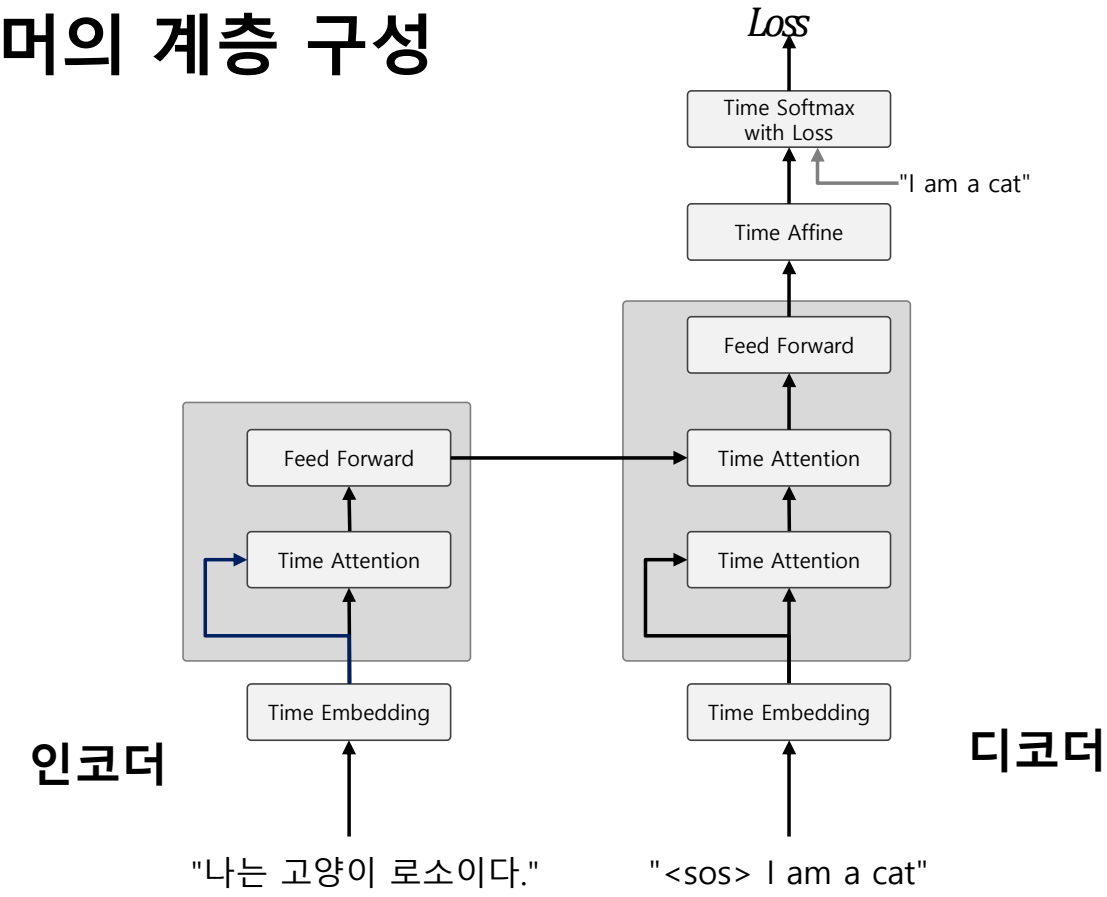
4. 포지셔널 인코딩(Positional Encoding) : Transformer 모델은 입력 시퀀스의 순서를 알 수 없기 때문에, 포지셔널 인코딩을 사용하여 각 단어의 위치 정보를 인코더와 디코더에 제공하여 순서의 의미를 학습할 수 있도록 한다.

Transformer 모델은 이후 다양한 변형 모델들(BERT, GPT 등)을 통해 발전을 거듭하며 자연어 처리 분야에서 표준이 되었으며, 기계 번역, 텍스트 생성, 문장 분류 등 여러 NLP 작업에 탁월한 성능을 보여준다.

알고리즘 설명 : https://sombio.tistory.com/26

1. 구글 Transformer 모델

트랜스포머의 계층 구성



2. 구글 BERT 모델

구글 BERT 모델

Google에서 2018년에 개발한 자연어 처리(NLP) 모델이다

BERT는 "Bidirectional Encoder Representations from Transformers"의 약자로, 양방향 인코더 표현을 사용하는 트랜스포머라는 의미이다.

기존의 NLP 모델들이 주로 문장의 앞에서 뒤로, 또는 뒤에서 앞으로 데이터를 처리한 것과 달리, BERT는 양방향으로 문맥을 이해할 수 있다. 즉, 문장 내에서 특정 단어가 앞뒤로 어떤 단어들과 연결되어 있는지를 동시에 고려하여, 보다 깊은 의미를 파악할 수 있다.

BERT의 구조

https://puppy-foot-it.tistory.com/370

BERT는 기본적으로 **트랜스포머 모델의 인코더(Encoder) 부분으로 구성된다.**

트랜스포머의 인코더는 셀프 어텐션(Self-Attention) 메커니즘을 사용하여 문맥을 이해하는데, BERT는 이 인코더를 여러 층으로 쌓아 문장을 깊이 있게 분석한다.

BERT의 구조는 크게 두 가지 버전으로 나뉜다:

- BERT-base: 12개의 인코더 층, 110M 파라미터
- BERT-large: 24개의 인코더 층, 340M 파라미터

주요 특징:

1. 양방향성(Bidirectionality): BERT는 문장을 양방향으로 처리한다. 즉, 단어의 의미를 이해할 때, 그 단어 앞과 뒤에 있는 단어들을 모두 고려한다. 이는 단어의 문맥을 더 깊이 이해하게 해준다. 기존의 단방향 모델들은 단어를 앞에서부터 순차적으로 읽으면서 학습하는데 반해, BERT는 문장의 모든 부분을 동시에 살펴본다.

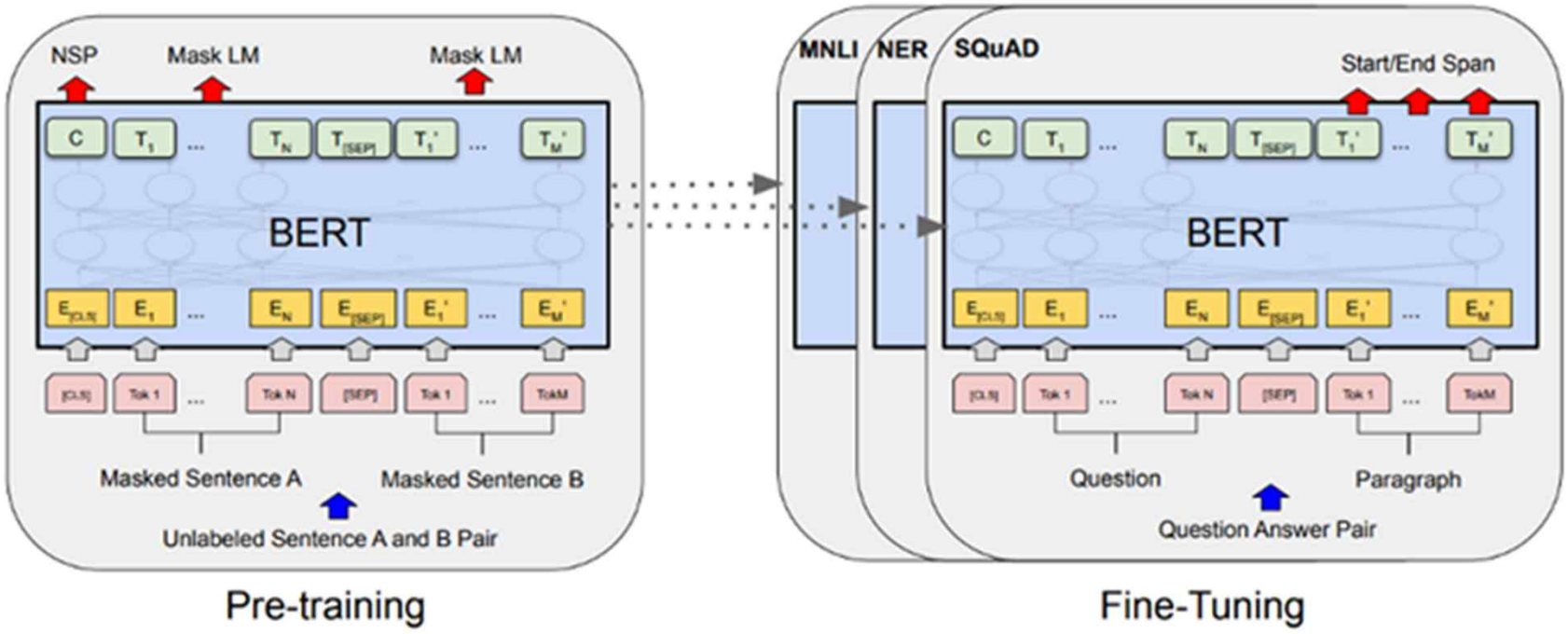
2. 사전 학습(Pre-training)과 미세 조정(Fine-tuning): BERT는 두 단계로 학습된다. 먼저, 대규모 텍스트 코퍼스를 사용하여 일반적인 언어 이해를 위한 사전 학습을 한다. 그 후, 특정한 NLP 작업(예: 문서 분류, 질의응답 등)에 맞춰 미세 조정을 한다. 이 방식은 다양한 작업에서 높은 성능을 나타낸다.

2. 구글 BERT 모델

3. Masked Language Modeling(MLM): BERT는 사전 학습 과정에서 입력 문장의 일부 단어를 "마스킹"하여 가린다. 그런 다음, 이 마스킹 된 단어들을 예측하도록 모델을 훈련시킨다. 이를 통해 모델이 단어의 문맥을 잘 이해할 수 있도록 돕는다.

4. Next Sentence Prediction(NSP): BERT는 두 문장이 연속적으로 이어지는지를 예측하는 작업도 수행한다. 이 과정을 통해 문장 간의 관계를 학습하고, 자연어 이해 능력을 더욱 향상시킨다.

BERT 모델 구조



https://happy-obok.tistory.com/23

BERT 모델 활용 분야

BERT는 다양한 NLP 작업에서 사용될 수 있다:

텍스트 분류: 이메일 스팸 필터링, 감정 분석 등

질의 응답: 질문에 대한 답변을 찾는 시스템

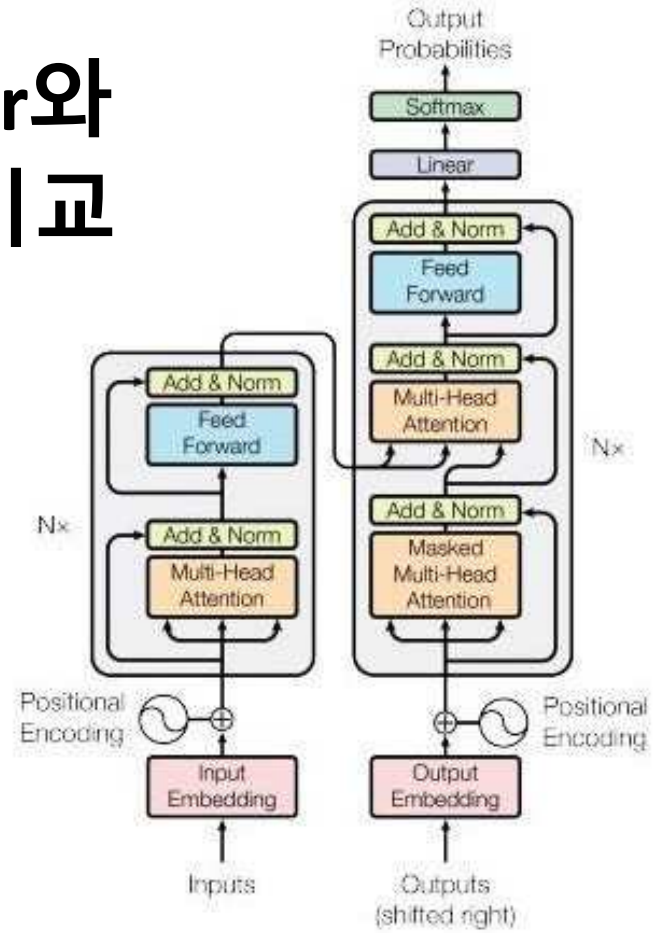
문서 요약: 긴 문서를 짧게 요약하는 작업

번역: 언어 간의 자동 번역

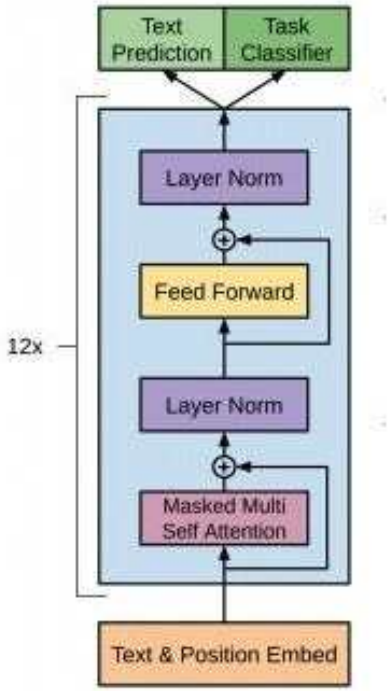
BERT는 등장 이후 많은 후속 모델들의 기반이 되었고, NLP 분야의 발전에 큰 기여를 했다. Hugging Face와 같은 플랫폼에서 쉽게 접근 가능하며, 다양한 언어 및 작업에 맞게 커스터마이즈 할 수 있다.

3. OpenAI GPT 모델

Transformer와 GPT 모델 비교



Transformer



GPT

OpenAI GPT 모델은

"Generative Pre-trained Transformer"의 약자로, 자연어 처리를 위한 인공지능 모델이다. 이 모델은 주어진 텍스트를 이해하고, 그 텍스트에 맞는 응답을 생성하는 능력을 가지고 있다.

GPT는 다음과 같은 주요 특징을 가지고 있다:

1. Transformer 아키텍처 :

GPT 모델은 "Transformer"라는 아키텍처를 기반으로 만들어졌다. 이 아키텍처는 순차적인 데이터를 처리하는 데 뛰어나며, 특히 자연어 처리 작업에 적합하다.

Transformer의 핵심은 "Self-Attention" 메커니즘이다. 이는 모델이 입력 텍스트의 모든 단어를 한 번에 살펴보면서 각 단어 사이의 관계를 이해할 수 있게 한다.

2. 사전 학습 (Pre-training)과 미세 조정 (Fine-tuning):

- **사전 학습 (Pre-training)**: 대규모 텍스트 데이터셋에서 모델을 먼저 학습시킨다. 이 과정에서 모델은 언어의 구조와 일반적인 문장 패턴을 학습한다.
- **미세 조정 (Fine-tuning)**: 특정 작업(예: 대화, 번역 등)에 맞게 사전 학습된 모델을 추가적으로 훈련시킨다. 이를 통해 모델이 주어진 작업에 더 최적화된다.

3. 자연어 생성 (Text Generation) :

GPT 모델은 사용자가 입력한 텍스트에 따라 자연스러운 응답을 생성하는 능력이 있다. 이를 통해 대화형 AI, 자동화된 글쓰기, 문서 요약 등 다양한 작업에 사용될 수 있다.

4. 언어의 맥락 이해 :

GPT는 주어진 문맥을 이해하고, 그에 맞게 다음 단어 또는 문장을 생성한다. 이는 긴 문장이나 문단에서도 일관된 답변을 제공하는 데 도움을 준다.

5. 다양한 응용 분야 :

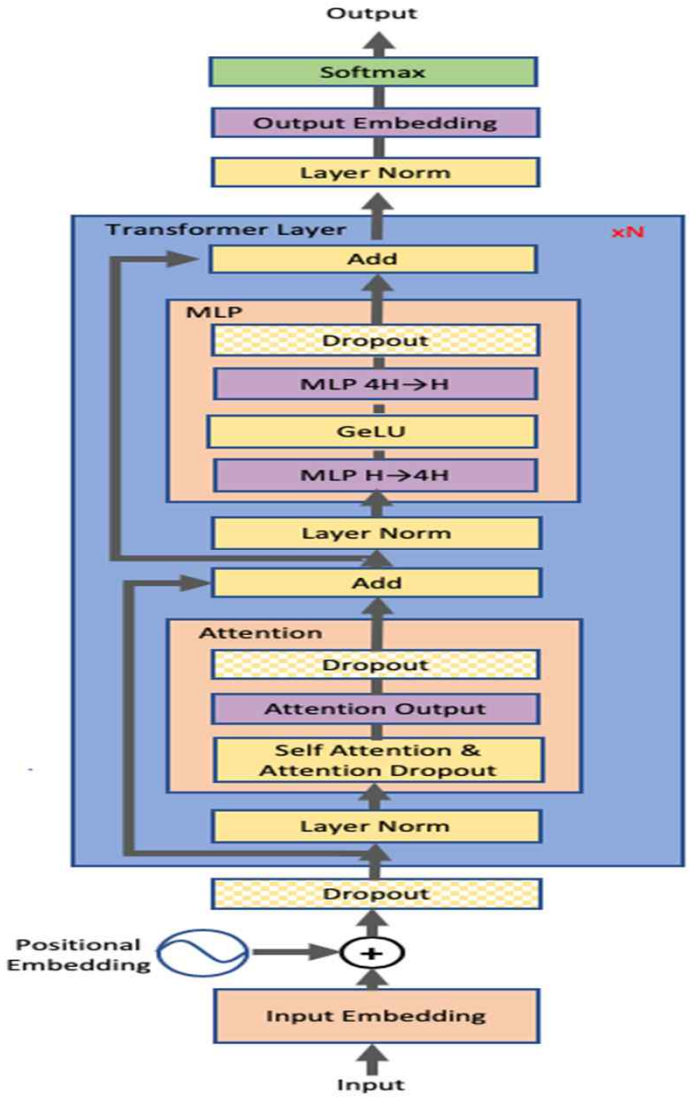
GPT 모델은 챗봇, 번역기, 텍스트 요약, 코드 생성, 창의적인 글쓰기 등 다양한 응용 분야에서 사용될 수 있다.

6. 버전 발전 :

GPT 모델은 계속해서 발전해 왔으며, 각각의 새로운 버전은 이전보다 더 큰 데이터셋과 더 많은 파라미터를 사용하여 훈련된다. 예를 들어, GPT-3는 1750억 개의 파라미터를 가진 매우 강력한 모델이다.

GPT 모델 구조

Transformer 의 Decoder 부분으로만 구성



GPT 모델의 발전 과정

GPT-1 (2018년)

최초의 GPT 모델로, Transformer 아키텍처를 기반으로 개발
1억 개의 파라미터를 가지고 있으며, 다양한 언어 모델링 태스크 수행 가능

GPT-2 (2019년)

GPT-1보다 훨씬 많은 15억 개의 파라미터를 가지고 있으며, 더욱 향상된 성능과 다양한 텍스트 생성 작업에서 놀라운 결과를 보여줌

3. OpenAI GPT 모델

GPT-3 (2020년)

1750억 개의 파라미터를 가지고 있으며, 인간 수준의 텍스트 생성 능력,다양한 분야의 지식을 습득하여 전문적인 글쓰기, 번역, 코딩 등 다양한 작업 수행 가능, 자연어 처리 분야의 새로운 지평을 열었으며, 다양한 산업 분야에 적용 가능성 제시

GPT-3.5 (2022년)

GPT-3의 후속 모델로, 더욱 향상된 성능과 새로운 기능 추가
챗GPT와 같은 대화형 모델에서 뛰어난 성능을 보임
코드 생성, 문제 풀이 등 다양한 작업 수행 능력 향상

3. OpenAI GPT 모델

GPT-4 (2023년)

GPT-3.5를 뛰어넘는 획기적인 성능 향상

다양한 전문 분야의 지식을 습득하여 더욱 정확하고 심층적인 답변 생성, 이미지 인식 능력 추가를 통해 텍스트와 이미지를 결합한 다양한 작업 수행 가능

GPT-4 Turbo (2024년 4월 9일)

한 번에 처리할 수 있는 단어량(token)을 128k로 증가]

GPT-4는 2021년 9월까지의 정보만 알고 있었지만, GPT-4 Turbo는 2023년 4월까지의 정보까지 학습했다. 기존 GPT-4 대비 평균 3분의 1 가격으로 사용이 가능하다.

GPT-4 Turbo with Vision 모델은 이미지를 분석할 수 있는 기능도 있다.

3. OpenAI GPT 모델

GPT-4o(2024년 5월 14일)

사람과 대등한 속도의 응답시간으로 실시간으로 대화가 가능하고, 중간에 사용자가 말을 끊어도 대화를 이어갈 수 있다.

사람과 영상 통화를 하듯이 대화를 할 수 있다.

사람의 말투와 표정을 읽고 감정을 이해할 수 있다.

이미지와 동영상을 실시간으로 인식하고 설명할 수 있다. 외모, 표정, 패션을 평가할 수 있다. 주위 상황을 통합적으로 인식하여 직업이 무엇인지, 사용자가 어떤 상황인지 판단할 수 있다.

글자를 인식하여 설명하고 수학 문제를 풀 수 있다.

적절하게 말투를 변화시키고, 웃거나 과장된 말투와 농담을 할 수 있다.

3. OpenAI GPT 모델

GPT-4o mini(2024년 7월 18일)

GPT-4o의 경량화 모델. 멀티모달 입출력을 지원한다.

OpenAI의 주장에 따르면 대형 모델인 Claude 3.5 Sonnet보다도 높은 성능을 보이며, 매우 빠른 속도를 자랑한다.

API 가격은 100만 토큰 당 입력 0.25달러, 출력 0.60달러로 성능이 더 낮은 모델인 GPT-3.5보다도 싸다.

GPT-4o mini가 공개됨에 따라 Chat GPT에서 GPT-3.5 모델은 더 이상 사용할 수 없게 되었으며 대신 그 자리를 GPT-4o mini가 차지하게 되었다.

https://namu.wiki/w/GPT-4

감사합니다