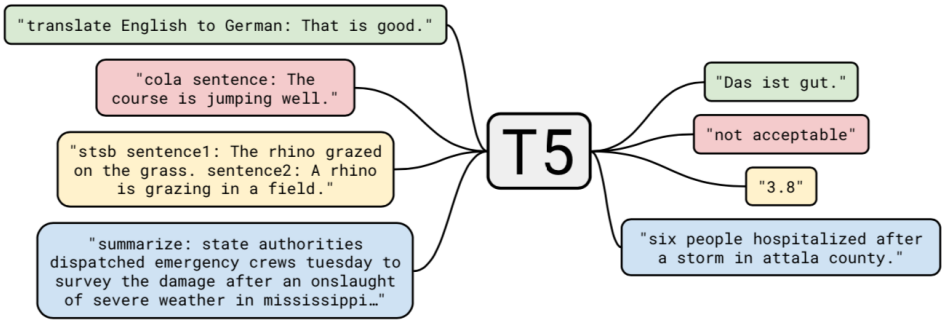
T5 모델(개념정리)

T5, Text-To-Text Transfer Transformer 모델



T5(Text-to-Text Transfer Transformer) 모델이 다양한 NLP 작업을 처리하는 방식을 설명하는 그림이다. T5는 **입력을 텍스트로 받아** 각 작업에 맞는 텍스트 출력을 생성하는 **텍스트-투-텍스트 모델**이다.

1. 영어에서 한국어 번역

- 입력: "영어에서 한국어 번역: That is good."
- 출력: "그거 좋다."

2. 문장의 적절성 평가

- 입력: "문장의 적절성 평가: 그 책은 걷는 것을 좋아함."
- 출력: "말이 안되는 문장"
- **설명**: T5는 문장의 적절성을 평가할 수 있는 모델이다. 여기서는 문장이 문법적으로 적절하지 않다고 판단한 결과이다.

3. 문장 간의 유사성 측정

- 입력: "문장 1: 나는 잠이 들었다, 문장 2: 나는 잠을 잤다."
- 출력: "9"
- **설명**: T5는 두 문장 간의 유사성을 숫자로 나타낼 수 있는 모델이다. 여기서 "9"는 두 문장이 매우 유사하다는 의미이다.

4. Self-supervised 학습

- 입력: 요약문에서 설명된 내용.
- 출력: "NLP 태스크에서의 전이학습 방법인 T5"
- **설명**: T5는 self-supervised 방식으로 학습된 모델이다. 대규모 텍스트 데이터를 활용해 레이블이 없는 상태에서 학습하며, 이후 다양한 NLP 작업에 적용할 수 있다.

결론

- T5는 **모든 NLP 작업을 텍스트-투-텍스트 문제로 통합**하여 처리하는 모델이다.
- T5는 번역, 문장 평가, 유사성 평가 등 다양한 NLP 태스크를 수행할 수 있는 강력한 모델이다.

∴ text 형태로 주어진 문제에서 text 정답 찾기인 것이다.

▼ <T5 모델에서 성능을 향상시키기 위해 사용된 방법론>

ⓐ **Model Architecture**: 기본적인 **Transformer 구조**가 **Encoder-only**나 **Decoder-only** 모델보다 더 높은 성능을 보인다. 즉, T5처럼 **Encoder-Decoder 구조**가 다양한 NLP 작업에서 더 좋은 결과를 낸다.

ⓑ **Pre-training Objectives**: **사전 학습**에서 노이즈가 섞인 데이터를 주고, 이를 **복구(denoising)** 하며 단어를 예측하는 방식이 더 효율적이다. 이는 모델이 데이터를 더 잘 학습할 수 있게 하고, 다양한 작업에 적용될 수 있는 일반적인 지식을 학습하게 한다.

ⓒ **Unlabeled Datasets**: 특정 도메인의 데이터를 사용하면 그 작업에 특화된 성능을 얻을 수 있지만, **데이터가 너무 적으면 과적합(overfitting)** 문제가 발생할 수 있다. 이는 모델이 특정 데이터에 너무 의존하여 일반화 성능이 떨어지는 상황을 말한다.

ⓓ **Training Strategies**: **Multitask learning**(다중 작업 학습)은 다양한 작업을 동시에 학습시키는 방식인데, 이 방식이 비지도 학습(unsupervised pretraining)과 비슷한 성능을 보인다. 학습할 때는 각 작업의 **비율을 조정**하여 적절히 학습해야 한다.

ⓔ **Scaling**: 모델의 크기(scale)를 늘리거나 **앙상블(ensemble)** 기법을 사용하면 성능이 더 향상될 수 있다. 특히, 작은 모델이더라도 **더 큰 데이터로 학습**하는 것이 성능 향상에 효과적이라는 점을 발견했다.

ⓕ **Pushing the Limits**: **110억 개의 파라미터**를 가진 대규모 모델을 훈련하여, 최신 최고 성능(State-of-the-Art, SOTA)을 달성했다. 또한, **1조 개 이상의 토큰**을 사용하여 대규모 학습을 진행하면서 모델 성능의 한계를 계속해서 밀어붙였다.

BERT와 T5의 차이점:

1. **하나의 token vs. 연속된 token에 대한 mask**

- **BERT**에서는 문장의 하나의 단어(token)만을 [MASK]로 변환한다. 즉, 문장의 일부 단어만 가리고, 그 가려진 단어를 맞추는 방식으로 학습된다.
- **T5**는 BERT와 달리, **연속된 여러 개의 단어**를 한꺼번에 [MASK]로 처리한다. 즉, 단일 단어가 아니라 **여러 개의 연속된 토큰**을 가리고, 이를 한 번에 맞추는 방식으로 학습한다. 예를 들어, "I love deep learning"이라는 문장에서 "deep learning"을 [MASK]로 처리하고, 이를 예측하게 하는 방식이다.

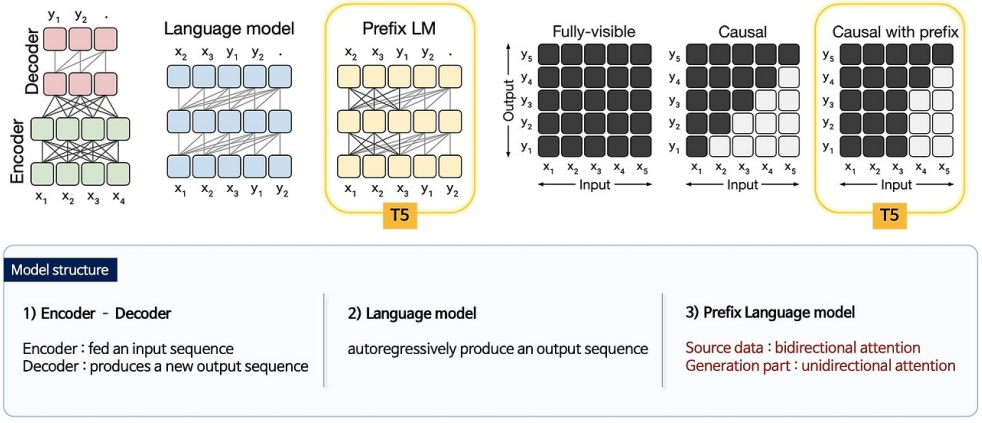
2. **Encoder-Decoder 구조**

- **T5는 Encoder-Decoder 구조**를 사용한다. 이는 입력(input)과 목표(target)를 따로 가지고 있는 구조다.
- **입력(input)**: T5 모델은 **[MASK]된 입력**을 받는다. 이 입력은 문장에서 일부가 [MASK]로 가려진 상태이다.
- **목표(target)**: 목표는 입력에서 가려진 [MASK] 부분을 정확히 예측하는 것이다. 즉, 모델이 [MASK]로 처리된 단어들을 복원하는 것이 목표다.

3. **Output Level에서 FFNN + Softmax로 Sequence 생성**

- T5는 출력 단계(Output Level)에서 FFNN(Feed-Forward Neural Network)와 **Softmax**를 통해 문장을 생성한다.
  - **FFNN**은 입력된 정보를 처리하여 최종적인 예측을 만들어내는 단계이다.
- **Softmax**는 각 단어에 대한 확률 분포를 계산하여, 가장 적합한 단어를 선택하게 한다. 이를 통해 [MASK]로 가려진 단어들이 무엇인지 예측하고, **연속된 시퀀스(sequence)를 생성**할 수 있다.

<정리>



Encoder-Decoder 구조 → T5

- **Encoder**는 입력 시퀀스(예: 텍스트)를 받아들이고, 이를 통해 문장의 의미를 이해하고 변환하는 역할을 한다.
- **Decoder**는 인코더의 출력 결과를 바탕으로 새로운 출력 시퀀스(예: 번역된 텍스트)를 생성한다.

Language Model (LM)

- **Autoregressive Language Model**은 시퀀스를 **순차적으로 생성**하는 모델이다. 즉, 첫 번째 단어를 생성하고 나면 그 단어를 기반으로 두 번째 단어를 생성하는 방식이다.
- GPT 모델이 대표적인 언어 모델이며, 문장을 순방향(왼쪽에서 오른쪽)으로만 생성한다.
- **Language Model**은 입력을 받으면 이를 바탕으로 순차적으로 출력 시퀀스를 만들어 낸다.

Prefix Language Model → T5

- **Prefix LM**은 입력 문장과 그에 따른 출력 문장을 함께 다루는 모델 구조다. 여기서는 입력 문장이 있고, 그 문장을 기반으로 **양방향 Attention**을 사용해 입력을 이해한 뒤, 그 입력을 기반으로 출력 시퀀스를 예측할 때는 **단방향 Attention**을 사용한다.
- 예를 들어, **T5** 모델은 입력 시 **양방향 Attention**을 사용하여 문장의 앞뒤 관계를 모두 이해하고, 출력을 생성할 때는 단방향(왼쪽에서 오른쪽)으로 단어를 생성한다.

Fully-visible, Causal, Causal with Prefix 모델

- **Fully-visible** 모델
    - 모든 입력을 한 번에 참조할 수 있는 모델로, 각 단어가 모든 다른 단어들을 볼 수 있는 구조이다. 문장의 모든 단어가 서로 영향을 주고받으며, 단방향이나 순차성이 없는 구조다.
- **Causal** 모델
    - **순차적**으로 한 단어씩만 참조하여 예측하는 구조다. 예를 들어, 현재 단어는 오직 그 이전 단어들만 참조하여 예측된다. GPT 같은 자기 회귀적 모델(autoregressive)이 여기에 해당된다.
- **Causal with Prefix** 모델→ **T5**
    - 입력에 대해 **양방향 Attention**을 사용하고, 생성 과정에서는 **단방향 Attention**을 사용하는 구조다. T5 모델이 이러한 방식을 사용한다. 입력을 분석할 때는 모든 단

어 간 관계를 고려하고, 문장을 생성할 때는 하나씩 순차적으로 단어를 생성하는 방식이다.