GPT 1

GPT1
GPT2
GPT3

Generative Pre-training

GPT 는 BERT와 다른 특성을 가지고 있다.

BERT가 transformer의 encoder부분을 잘라서 사용한다면, GPT는 transformer의 decoder 부분을 잘라서 사용한다.

decoder 부분 맨 아래에 있는 masked self-attention 부분 때문인데 masked self -attention은 input에 대해서 각 input이 전의 것만 참조 할 수 있도록 만든 것이다. 예를 들어, i am a boy 가 input으로 들어왔다고 가정하면 am은 참조할 때 i 밖에 참조를 못하는 것이다.

이러한 부분 때문에 GPT는 전반적으로 bidirectionality가 필요한 task에서 안좋은 성능을 가진다. 예를 들어 빈칸채우기나 전체 글을 읽고 짧은 답을 내는 task가 있다.

GPT1

GPT1에서는 다양한 special token을 활용해, fine-tuning의 성능을 극대화 시킨 모델이다. 문장 끝에 넣어준 토큰을 활용해 원래 원하는 downstream task의 query 벡터로 활용 된다. 같은 transformer 구조를 별도의 학습없이 여러 task에서 활용할 수 있다는 장점이 있다. 같은 맥락으로 수행하고자 하는 task에 대한 데이터가 얼마 없을때 pre-training 된 데이터를 fine tuning 부분에 자연스럽게 전이 학습 시킬 수 있다.

GPT2

GPT2

GPT3

GPT3