[7주차] GPT2

📌 **GPT-2 - Language Models are Unsupervised Multitask Learners**

Introduction

기계학습은 큰 dataset과 고용량의 모델, 지도학습 등을 통해 빠르게 발전해왔습니다. 현재 기계학습을 개발하는 주된 방법은 목표 과제에 맞는 dataset을 찾아서, 이를 학습/검증 단계로 나누어 학습 후 IID(independet and identically distributed)로 성능을 측정하는 방법입니다. 이는 좁은 범위의 과제에서는 매우 효과적이나 범용적인 이해를 필요로 하는 독해나 다양한 이미지 분류시스템 등의 문제에서는 높은 성능을 내지 못했습니다.

이러한 방법으로 개발된 모델들은 사소한 변경에도 정확도가 떨어지기 쉽고 지도학습을 사용해 매우 좁은 범위의 문제에서만 뛰어난 능력을 발휘합니다. 그래서 **데이터를 수동 분류하는 과정 없이도 더 범용적인 모델을 개발**할 필요가 있습니다.

Approach

모델 접근법의 핵심은 **Language Modeling**입니다.

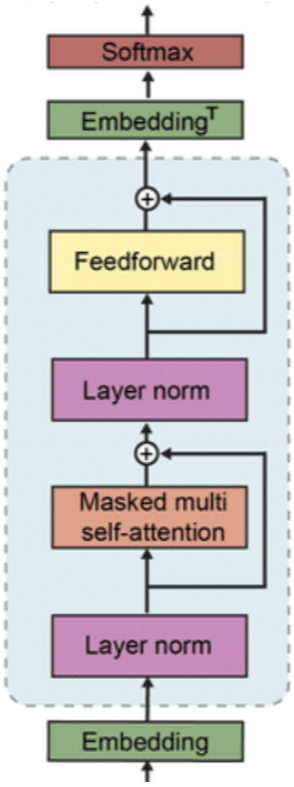
$$p(x) = \prod_{i=1}^{n} p(s_n \mid s_1, \ldots, s_{n-1})$$

self-attention 구조를 갖는 Transformer가 이런 조건부 확률을 잘 계산합니다. 하나의 task를 수행하는 프레임워크는 p(output | input)으로 표현될 수 있습니다. multitask를 수행하는 프레임워크는 task에 대한 정보도 필요로 하기 때문에 p(output | input,task)를 모델링해야 합니다. 특정 task를 설정하는 것은 architecture 레벨에서 수행되기도 하지만 그러나

McCann et al.(2018)에서 알 수 있듯이 **언어는 과제/입력/출력 모두를 일련의 symbol로 명시**하는 유연한 방법을 제공합니다.

- 예를 들어 번역학습은 (프랑스어로 번역, 영어 텍스트, 프랑스어 텍스트)로 표현된다 (translate to french, english text, french text).
- 독해는 (질문에 대답, 문서, 질문, 대답)이다(answer the question, document, question, answer).

Model



GPT-2는 논문 Attention Is All You Need에서 제시한 트랜스포머 구조에서 인코더 블록을 제거하고 디코더 블록만 사용한 모델입니다.

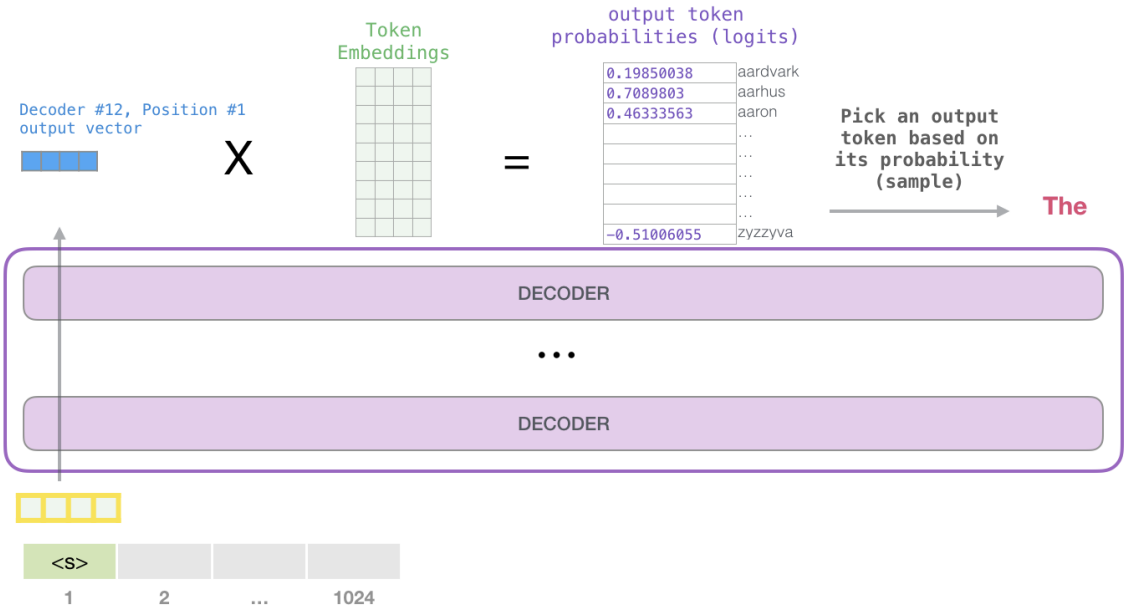
1. 각 단어의 순서들을 고려한 임베딩 행렬을 입력합니다.
2. 각 디코더의 Self Attention과정을 거친 후 신경망 레이어를 통과합니다.
3. 출력값을 임베딩 벡터와 곱해줍니다. 그 결과 값은 각 단어가 다음 단어로 등장할 확률값으로 출력됩니다.
4. 이 중 가장 높은 확률값을 가지는 단어를 출력해 이값은 곧 다음 입력값이 됩니다.

GPT-1과 동일하게 **Masked Self-Attention**, **Feed Forward Neural Network**를 포함하는 Transformer의 Decoder Block을 여러개 쌓아 구성합니다. Decoder 모듈들은 모두 같은 구조를 갖고 있지만, 모듈별로 **개별적인 가중치**를 갖고 다음 단어를 예측하기 위한 연산을 수행합니다.

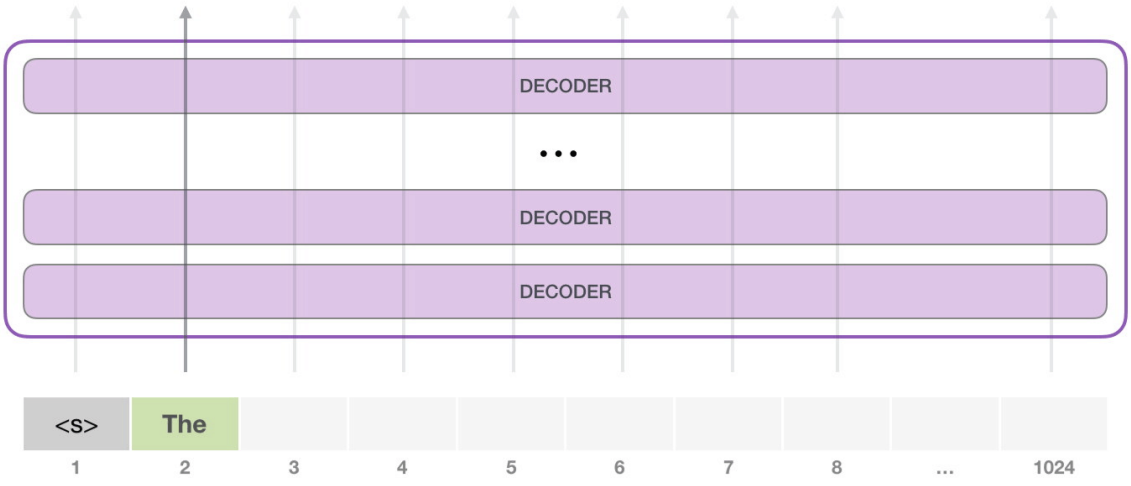
Input으로는 단어를 **Byte Pair Encoding** 방식으로 분해해 토큰으로 나눈 뒤, 위치에 대한 정보를 담고있는 **positional encoding**을 더해 입력합니다.

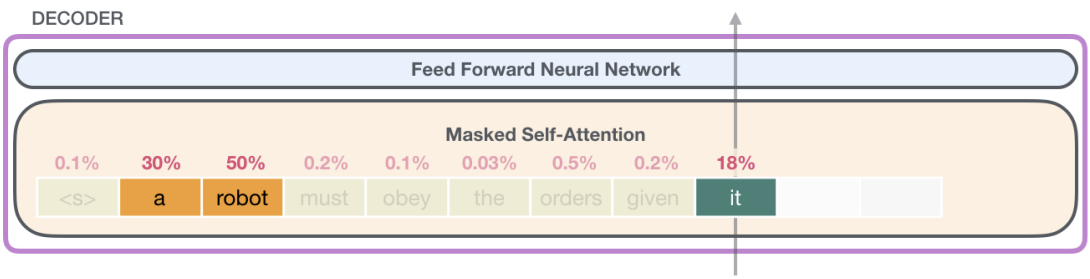
Byte Pair Encoding (BPE)

gpt-2는 Byte Pair Encoding를 거친 토큰을 입력 단위로 사용합니다. **BPE는 서브워드를 분리하는 알고리즘**으로, 빈도수에 따라 문자를 병합하여 서브워드를 구성합니다. 단어를 문자(char) 단위로 쪼갠 뒤, 가장 빈도수가 높은 쌍을 하나로 통합하는 과정을 반복하여 토큰 딕셔너리를 만듭니다. 이를 통해 13만개의 Token 사전을 256개의 Token 사전으로 줄일 수 있었습니다. 또한 이는 모델이 **어떤 종류의 전처리, 토큰화, 단어 사전 크기라도 모두 사용할 수 있게 합니다.**



input된 벡터는 아래와 같이 Decoder 블럭을 직렬로 통과한 뒤 **token embedding**과 곱해져 가장 높은 확률의 단어로 출력됩니다.



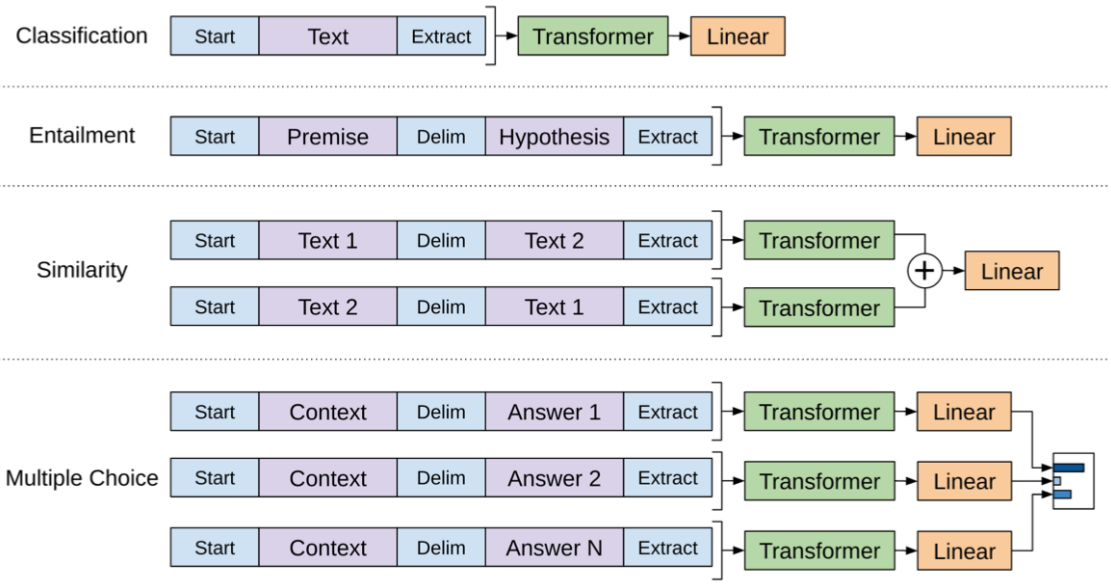


눈에 띄는 점은, Window Size가 1024개로 설정되어있어서 Decoder에 1024개의 레일을 통해 각각의 토큰들이 Decoder를 통과하는 길이 정해져있다는 점입니다. 이를 통해, 한 문장이 쭉 입력되고 있어도 이전 토큰 데이터들이 유지되기 때문에 **Attention을 통해 문장 이전 토큰들에 대한 정보를 가져올 수 있습니다.**

GPT-1 VS GPT-2

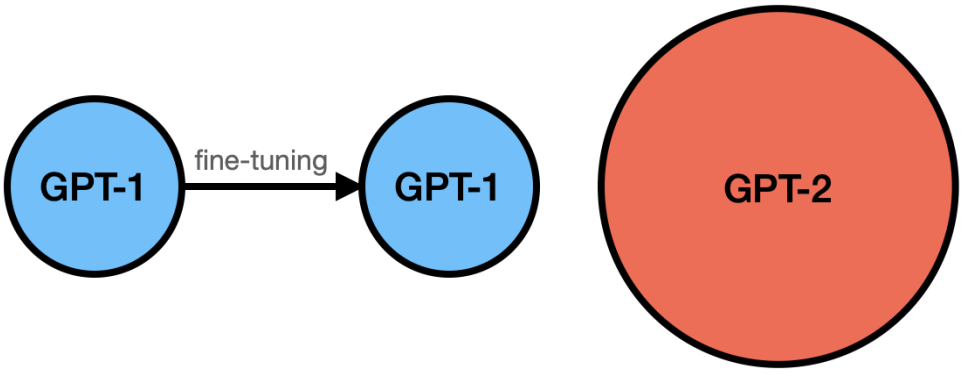
GPT-2가 GPT-1과 다른점은, fine-tuning 단계(Supervised Learning 단계)를 거치지 않는다는 것입니다. GPT-2는 특정한 task에 특화되도록 따로 학습을 진행시키지 않고, **Unsupervised Learning 만으로 Multitask learning을 실현시켜 모든 task에 대응**할 수 있도록 하였습니다. task의 구분을 위해서는 GPT-1 에서 task에 대해 input 정보를 수정했던 것처럼 input 정보에 특정한 token을 붙여서 task 정보를 함께 input으로 전달해 처리하였고, 모델은 해당 token을 확인하고 적절하게 task를 수행합니다.

▼ **GPT-1의 입력양식**



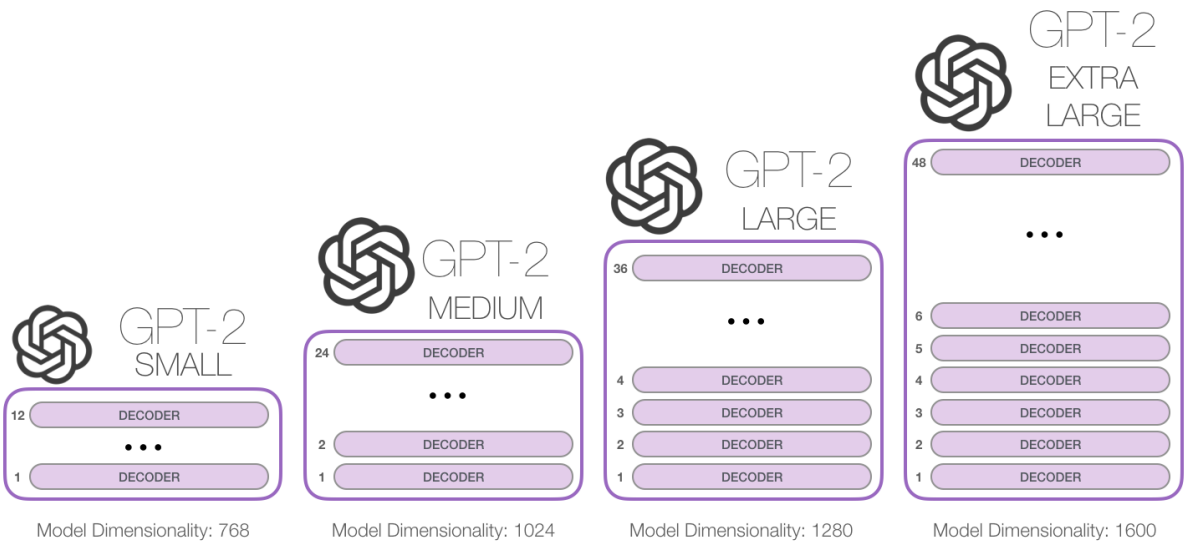
GPT-1의 input 양식

특정한 task가 아닌 모든 task에 대해서 범용적으로 작동하는 모델인 만큼, **GPT-2**는 GPT-1보다 훨씬 많은 데이터를 필요로 하고, 모델의 크기 또한 파라미터의 개수도 최대 11배 이상이며 레이어도 최대 4배 많은 개수를 갖습니다. 단어 사전의 크기 역시 50,000개의 단어로 늘어났고, context size 또한 1024 token으로 2배로 늘어났습니다. 한 번에 처리할 수 있는 token의 개수는 무려 8배나 많아졌습니다.



본 논문에서는 GPT-2 모델에서 성능을 향상시키기 위해 디코더 모듈의 개수를 늘리는 방법을 사용하였습니다. **디코더의 수가 많아질수록 연산량은 많아지지만, 좀 더 정교하게 문장을 생성하고 task를 처리할 수 있음을 아래에 기술할 실험을 통해 밝혀냈습니다**. GPT-2의 발표와 함께 대중들에게 공개되었던 모델은 GPT-2 SMALL 모델이며, 비슷한 시기에 등장한 구글의 NLP 모델인 BERT LARGE 모델은 GPT-2 MEDIUM 모델과 비슷한 크기(파라미터 수)를 보여줍니다. 따라서, 해당 시기에 GPT-2 Extra Large 모델은 엄청난 크기의 모델이였음을 짐작해볼 수 있습니다.

아래에서 GPT-2 Small이 GPT-1와 비슷한 크기이며, **GPT-2 Medium이 BERT와 비슷한 크기**입니다. 일반적으로 GPT-2라 하면 Extra Large 모델을 의미합니다.



Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Results

GPT-2 모델이 주어진 input 문장을 가지고 새로운 기사를 작성한 예시입니다. 영어를 완벽하게 구사하는 유니콘이 안데스 산맥에서 발견되었다는 내용인데, GPT-2 모델이 정말 그럴듯하게 글을 작성했습니다.

🤖 **Context (human-written):** In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is fi- nally solved.

Dr. Jorge Pe´rez, an evolutionary biologist from the University of La Paz, and several com- panions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pe´rez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pe´rez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pe´rez.

Pe´rez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pe´rez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pe´rez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pe´rez, "In South America, such incidents seem to be quite common."

> However, Pe´rez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. “But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,” said the scientist.

Conclusion

OpenAI는 **학습 과정에서 Labeled Data를 필요로 하지 않으면**서도 다양한 task에 적용할 수 있는 **범용적인 모델인 GPT-2**를 제시하였습니다. 학습 데이터와 모델의 크기를 훨씬 키워 Multitask Learning의 효과를 통해 동일한 모델이 여러 task에 적용될 수 있도록 하였습니다.

여전히 요약과 같은 일부 task에서는 가장 좋은 성능을 내지는 못하고 있었지만, fine-tuning된 다른 SOTA 모델들과 비교했을 때 준수한 수준의 성능 혹은 더 우수한 성능을 보여주면서 **Unsupervised Learning을 통한 모델 학습의 가능성**을 보여주었습니다.

KoGPT2

KoGPT2는 SKT-AI에서 GPT-2의 한국어 성능 개선을 위해 개발한 모델입니다. GPT-2는 주어진 텍스트의 다음 단어를 잘 예측할 수 있도록 학습된 언어모델이며 문장 생성에 최적화되어 있습니다. **KoGPT2는 부족한 한국어 성능을 극복하기 위해 40GB 이상의 텍스트로 학습된 한국어 디코더decoder 언어모델**입니다.

KoGPT2에서 사용된 tokenizer는 CBPE(Character Byte Pair Encoding)가 사용되었고, 대화에 자주 쓰이는 이모티콘, 이모지등을 추가하여 토큰의 인식 능력을 높였다고 합니다.

한국어 위키 백과 이외, 뉴스, 모두의 말뭉치 v1.0, 청와대 국민청원 등의 다양한 데이터가 모델 학습에 사용되었습니다.

> https://github.com/SKT-AI/KoGPT2

Refenrences

https://greeksharifa.github.io/nlp(natural language processing) / rnns/2019/08/28/OpenAI-GPT-2-Language-Models-are-Unsupervised-Multitask-Learners/

https://velog.io/@alsbmj012123/논문리뷰Language-Models-are-Unsupervised-Multitask-Learners

https://jalammar.github.io/illustrated-gpt2/