
BERT

GPT-1, 2, 3

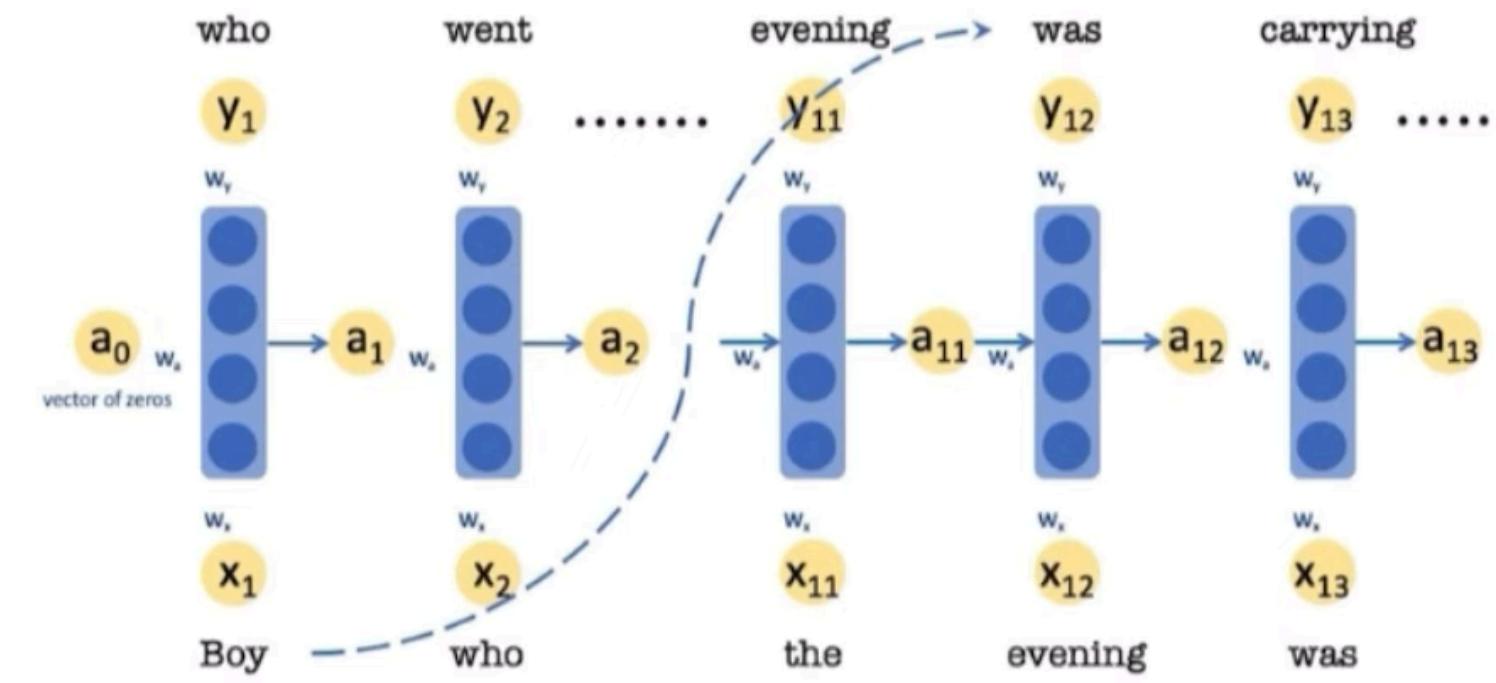
팀 레모네이드: 장우진 김선준 장재우 강희준 박은수

2025.10.13

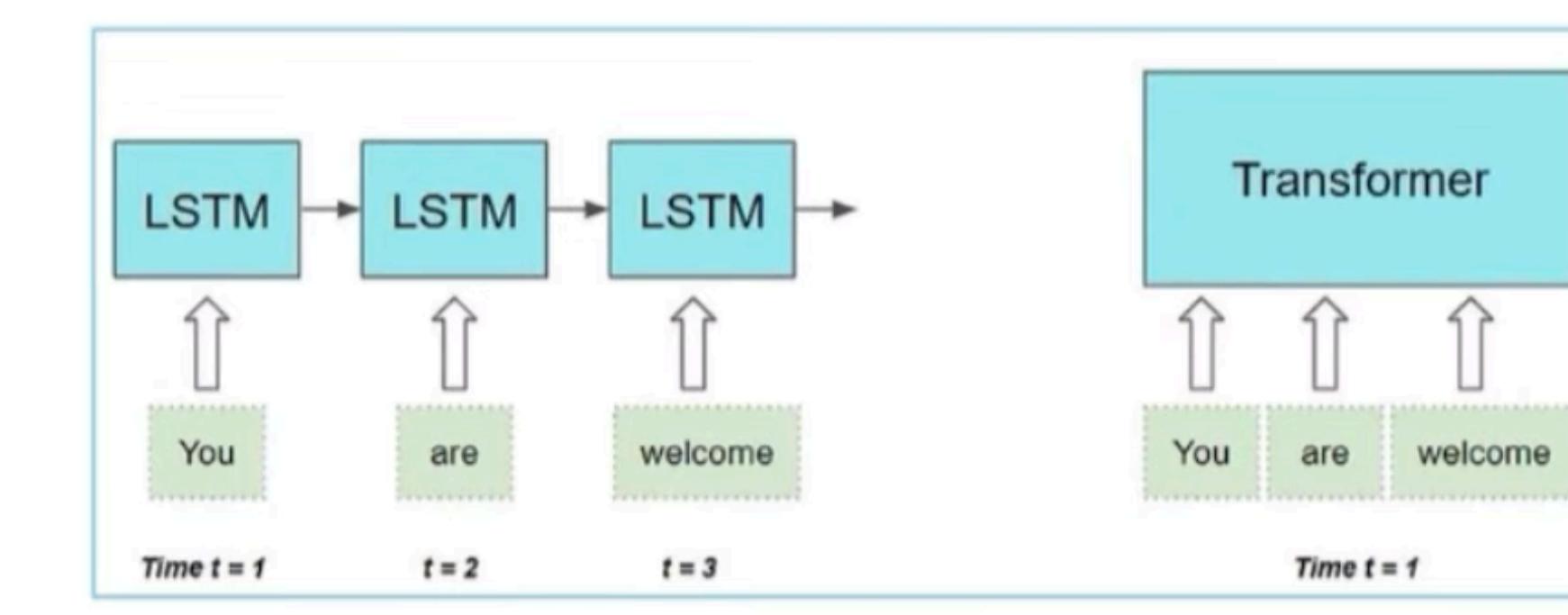
개요

Trasformer	4 page
BERT	7 page
GPT-1	24 page
GPT-2	36 page
GPT-3	45 page
종합 비교	63 page
참고문헌	64 page

previously on NLP



RNN



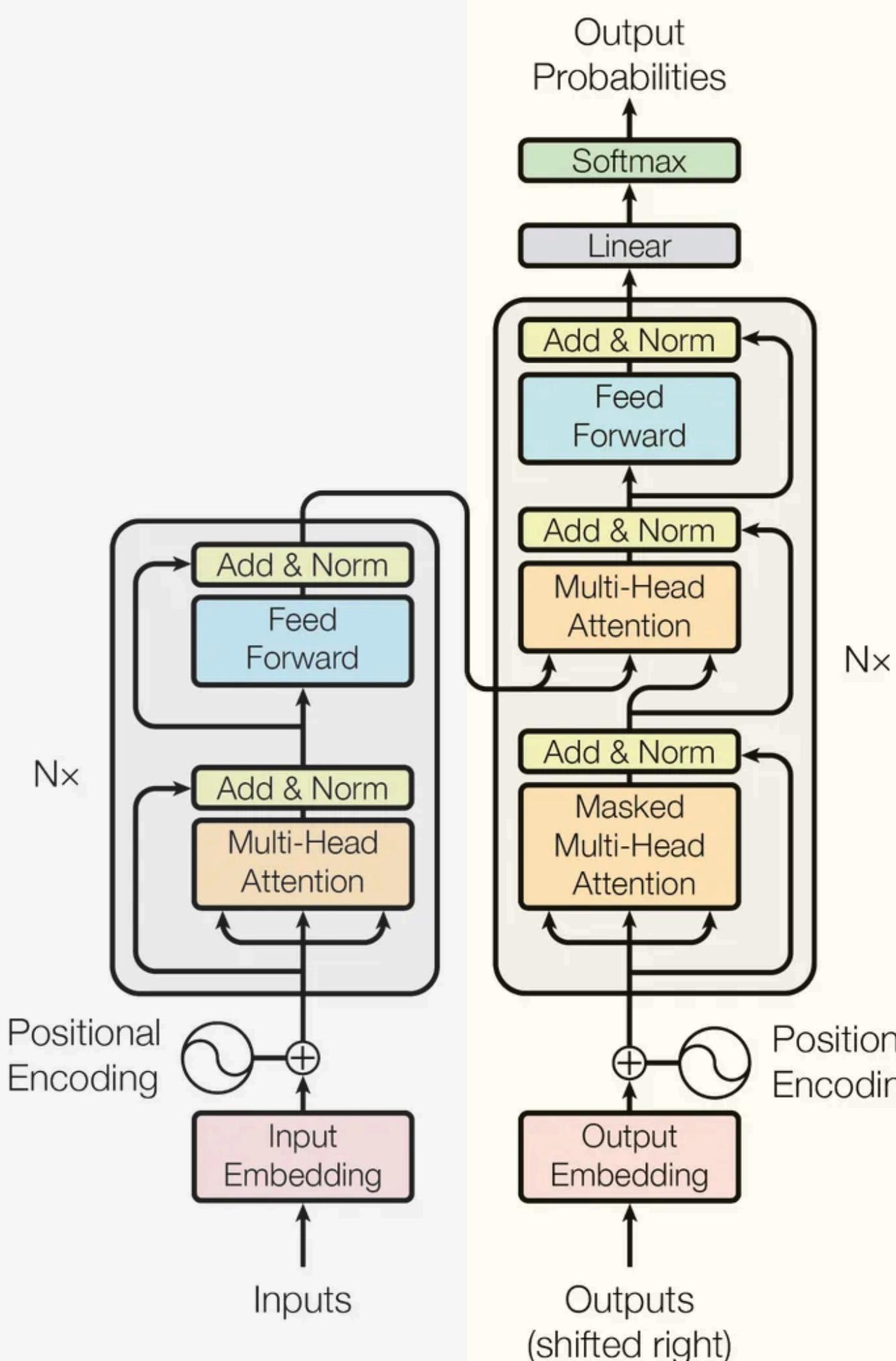
LSTM

시퀀스 예측을 위한 첫 시도들
sequential processing이기 때문에 느림
짧은 문장에는 잘 작동하지만 긴 문장에는 힘들어함
(장거리 의존성, 단어가 30개 이상이면..?)

Transformer

Encoder

입력 문맥을 깊이 있게
이해하고
정보를 압축하여
디코더에게 전달



Decoder

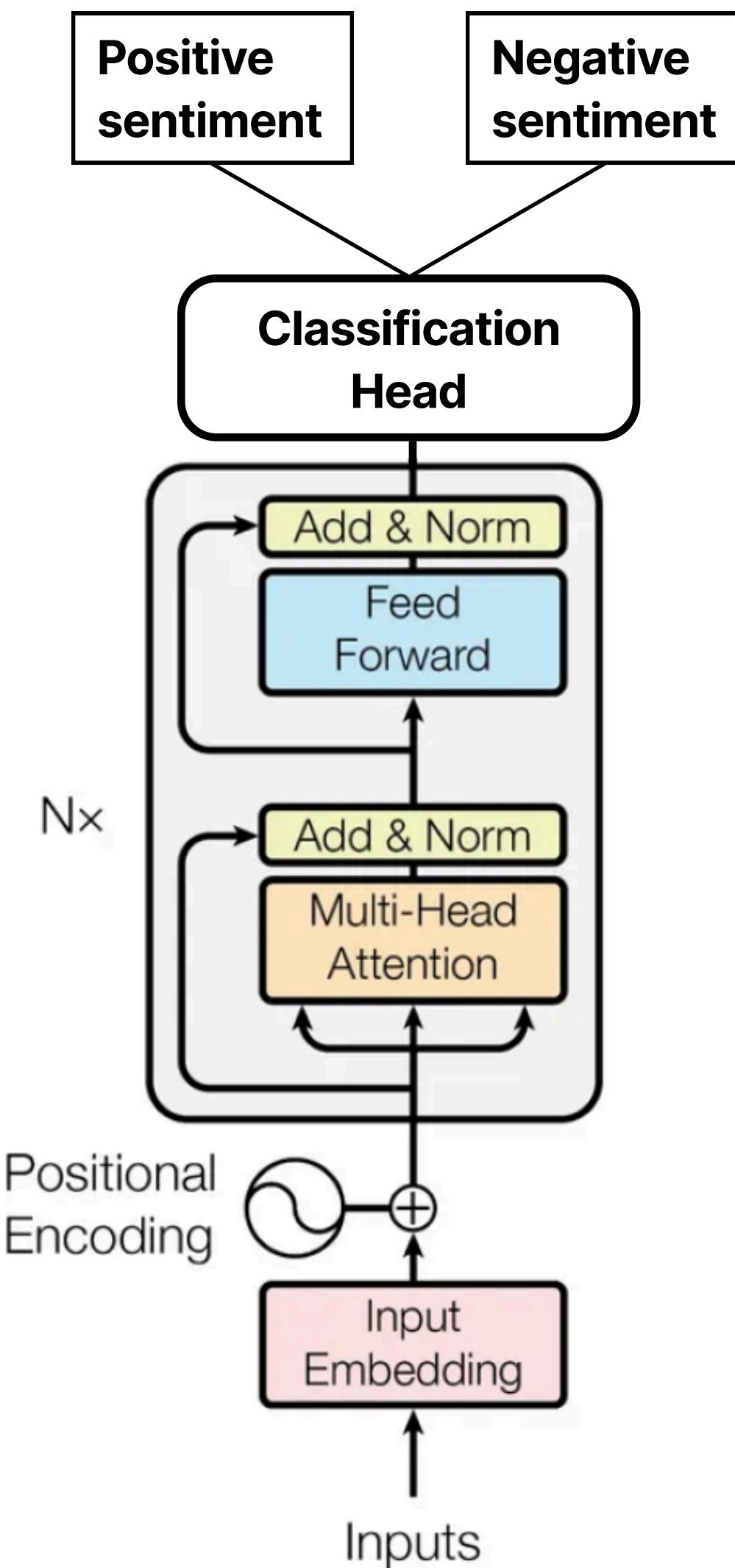
인코더에서 받은 정보를
바탕으로 다음 단어를
순차적으로 예측하고
출력하여 최종 결과 문장을 생성

Transformer

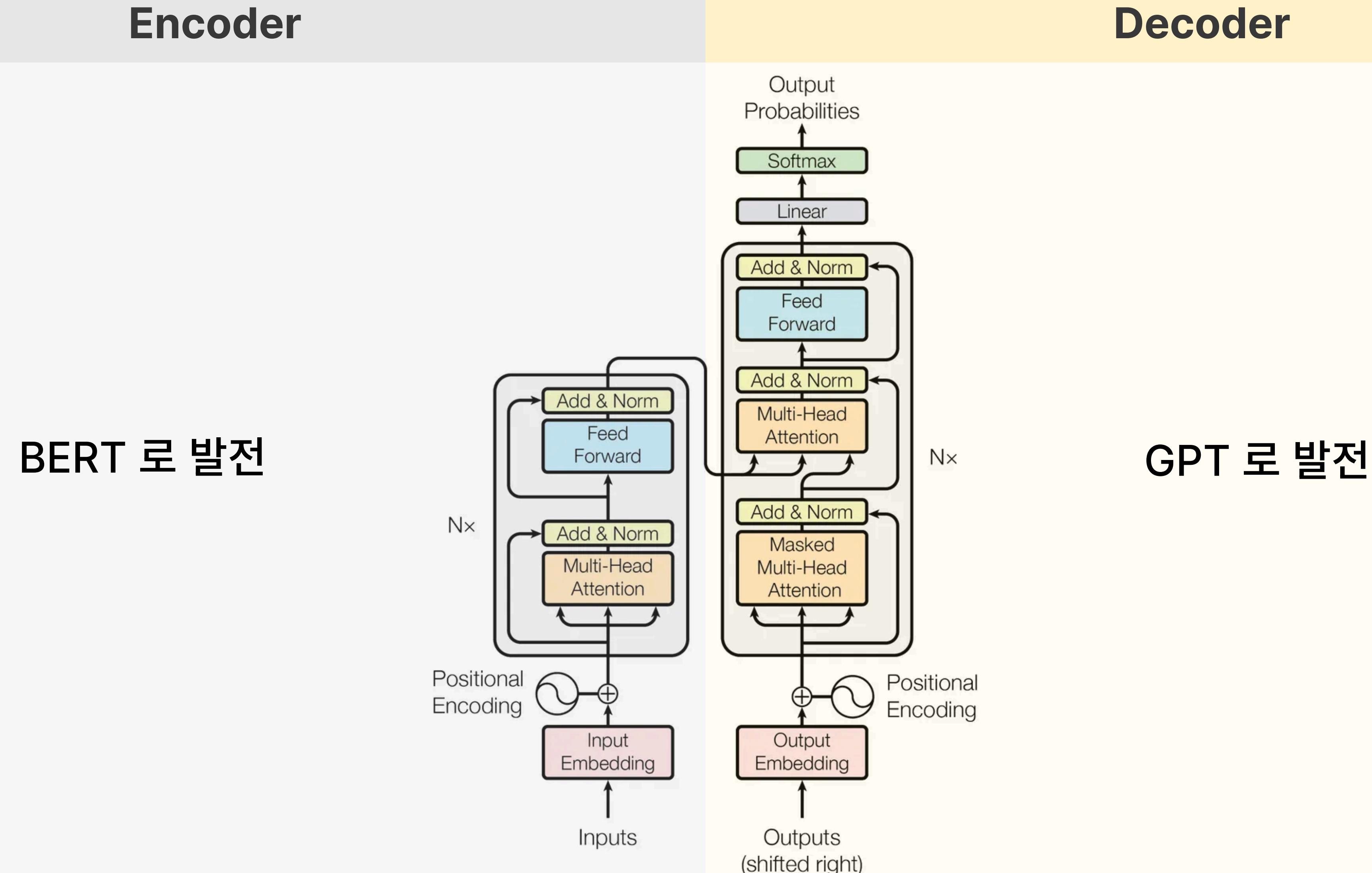
유연한 응용 가능

Sentiment classification

텍스트의 감정을 분석하고 분류
encode 위에 task에 맞는 레이어만 붙이면 됨
글자 생성 필요없음 → decoder 필요없음



Transformer



BERT

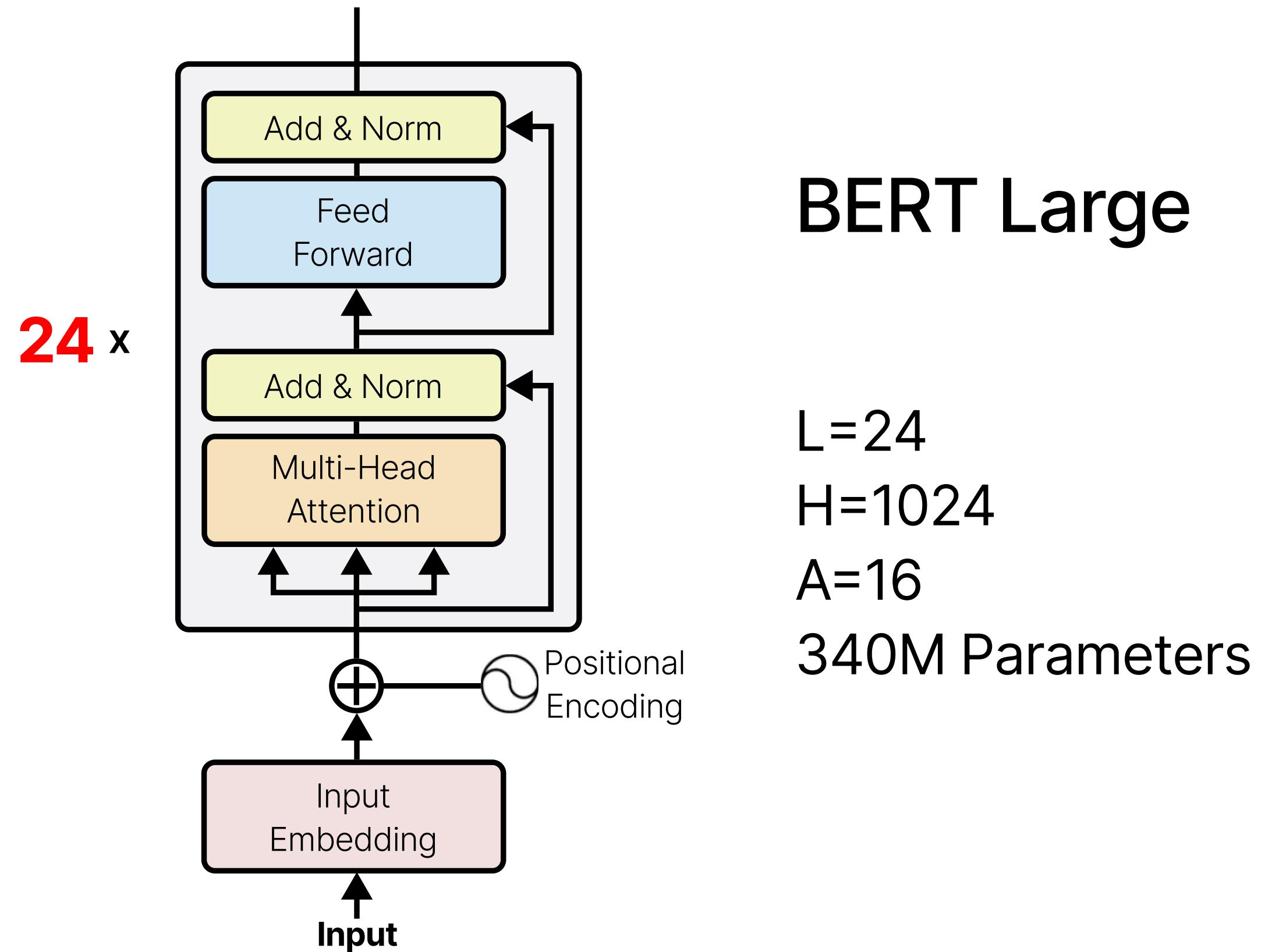
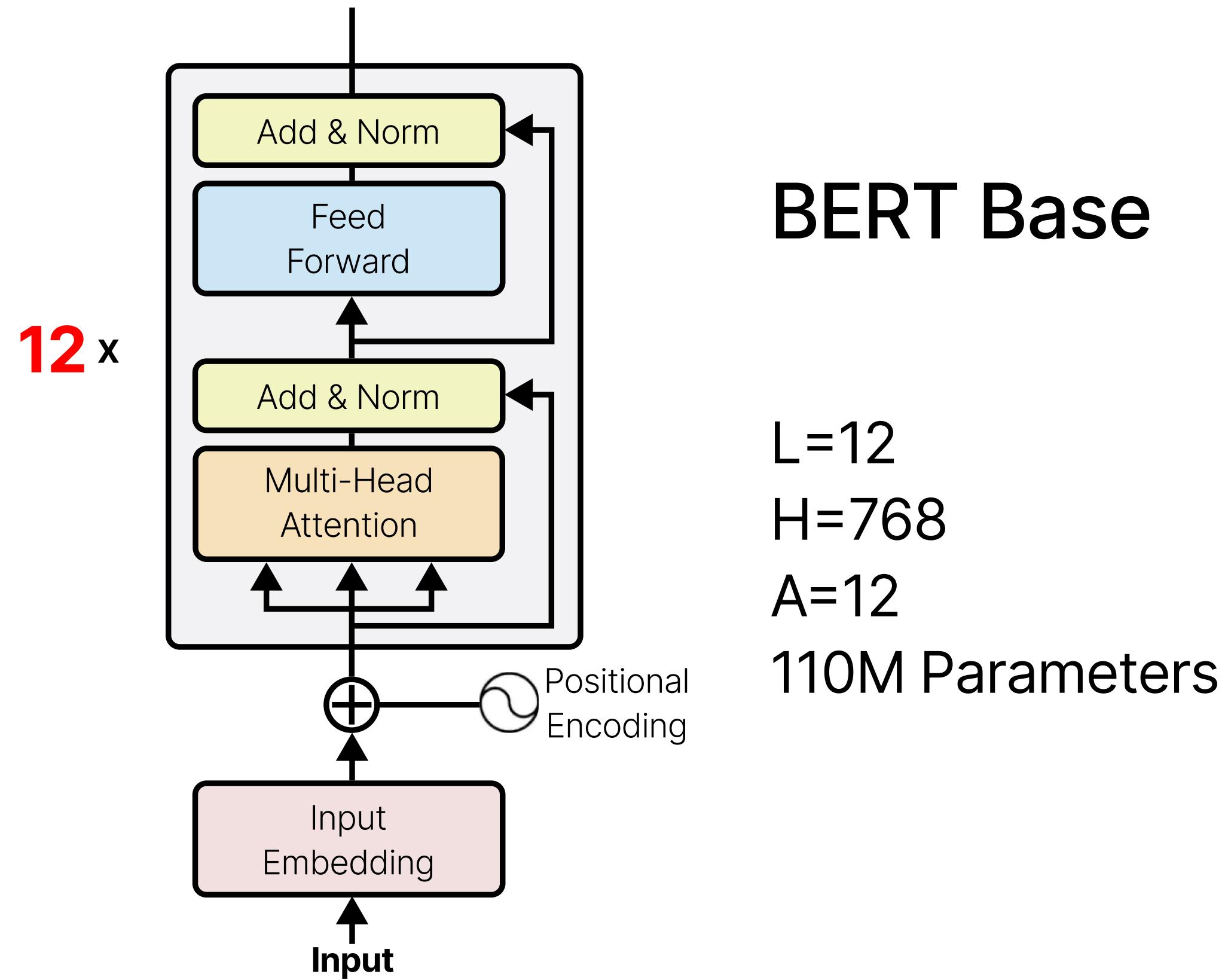
Encoder

Bidirectional
Encoder
Representations from
Transformers



BERT

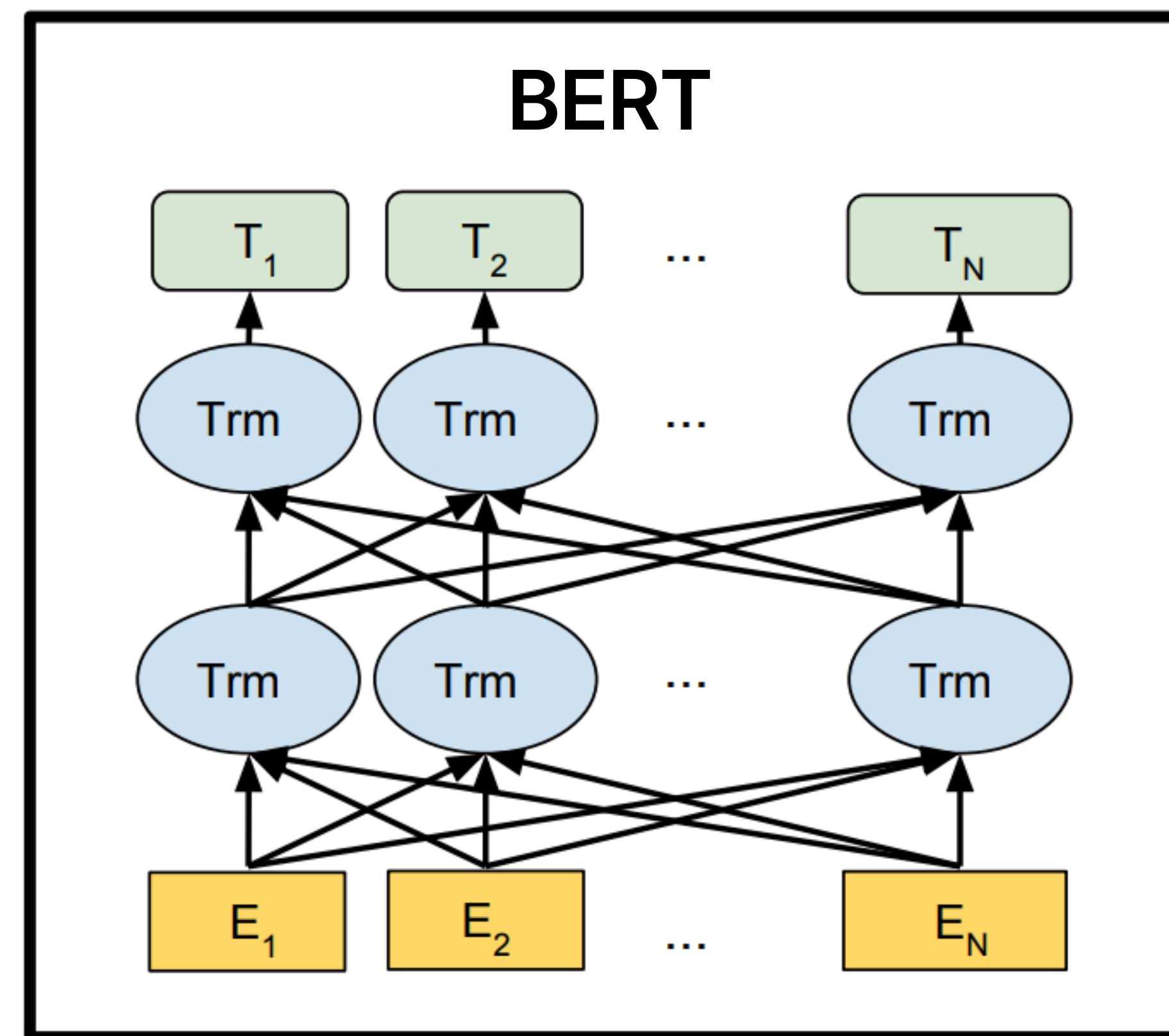
Architecture



L = 블록 수, H = 하든 노드 사이즈, A = self attention heads 개수

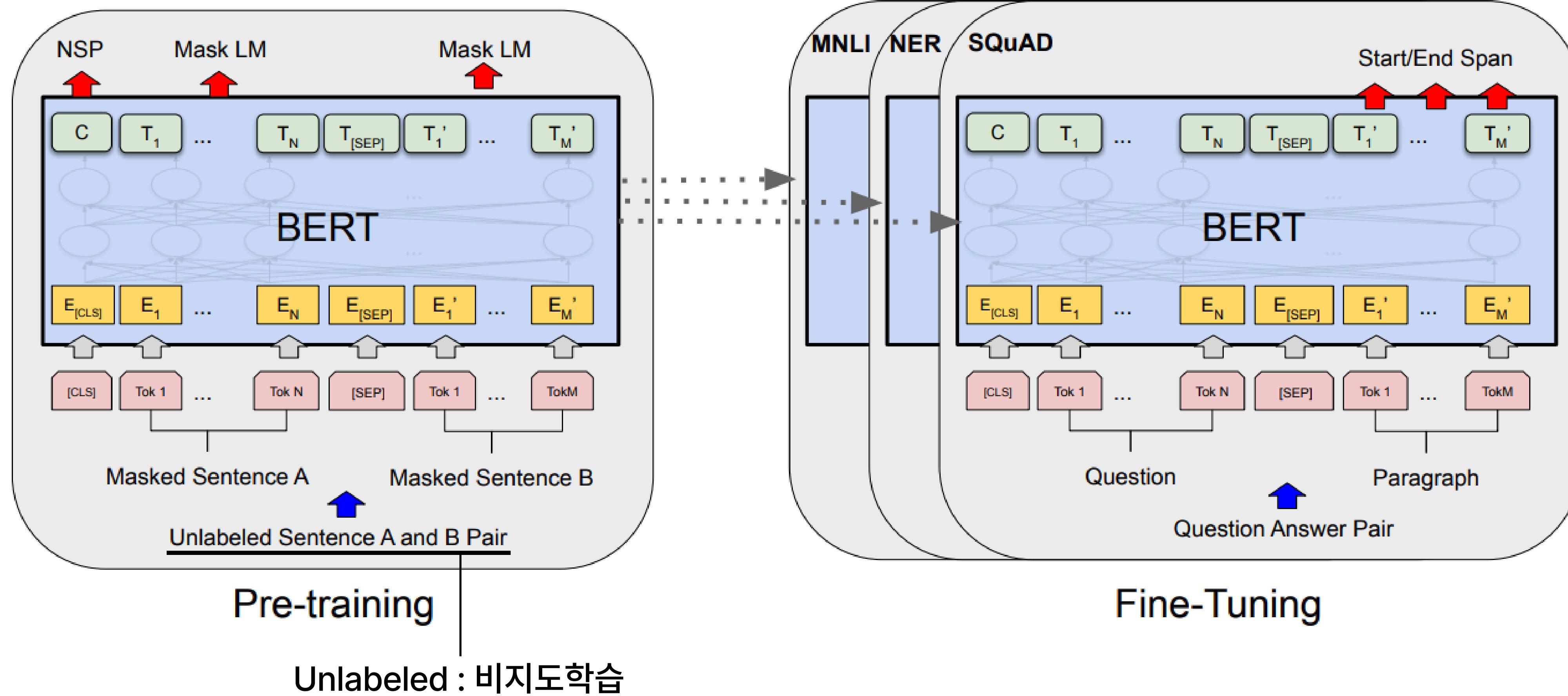
BERT

Architecture



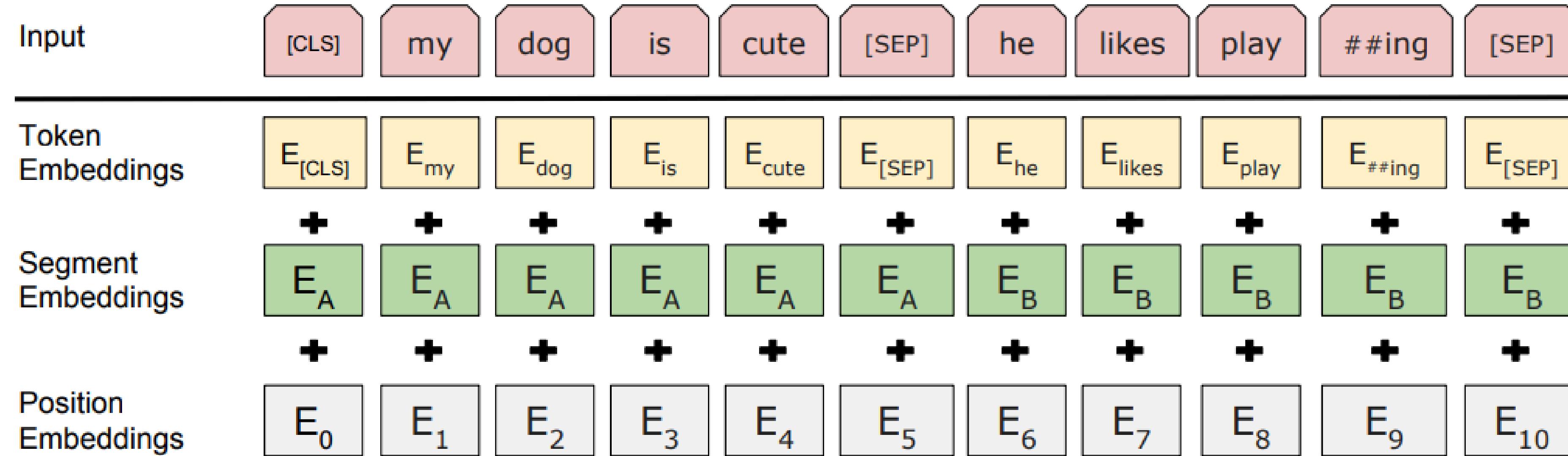
BERT

training 순서



BERT

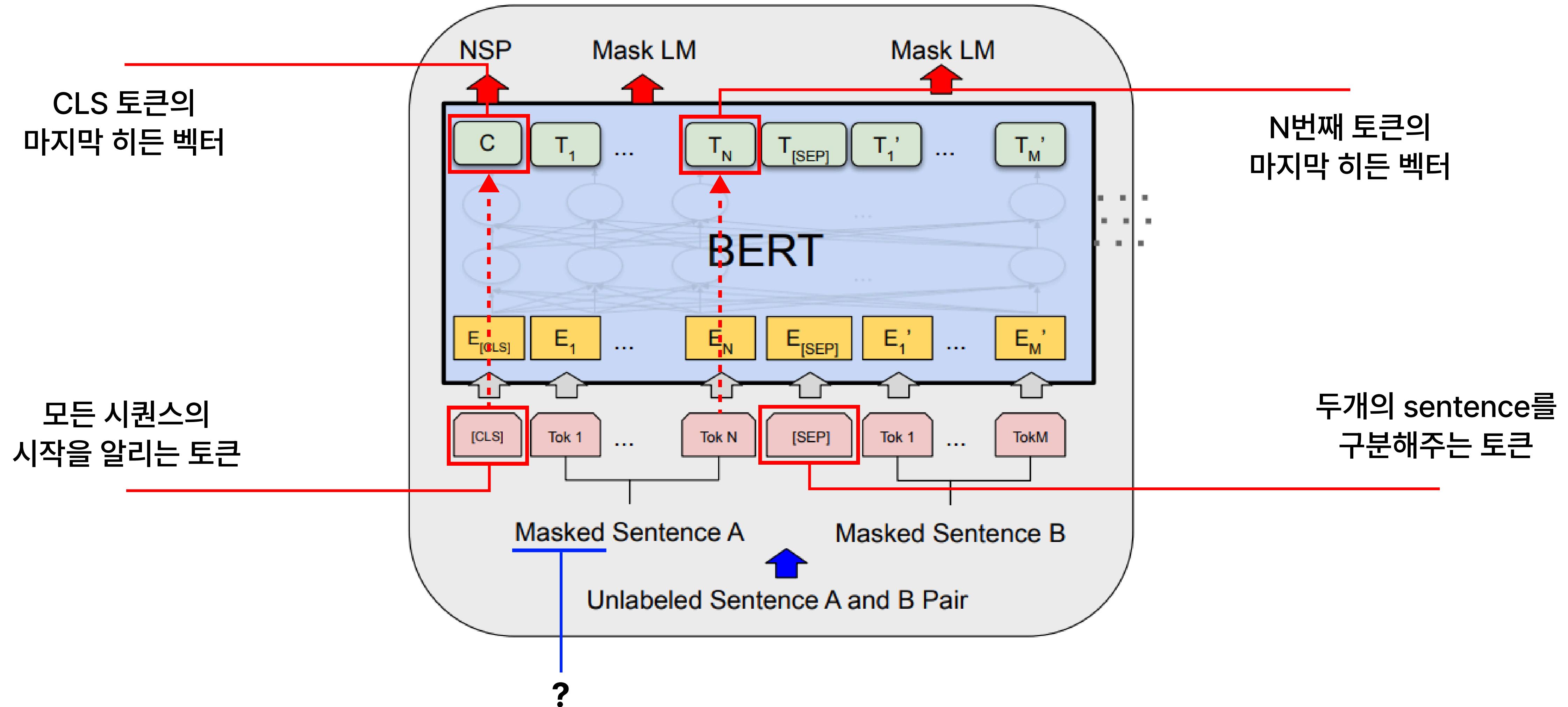
입력 표현



Token embedding : WordPiece를 사용해 3만개의 단어 조각으로 나눔
특별 토큰들도 여기에 포함됨
[CLS], [SEP], [MASK]

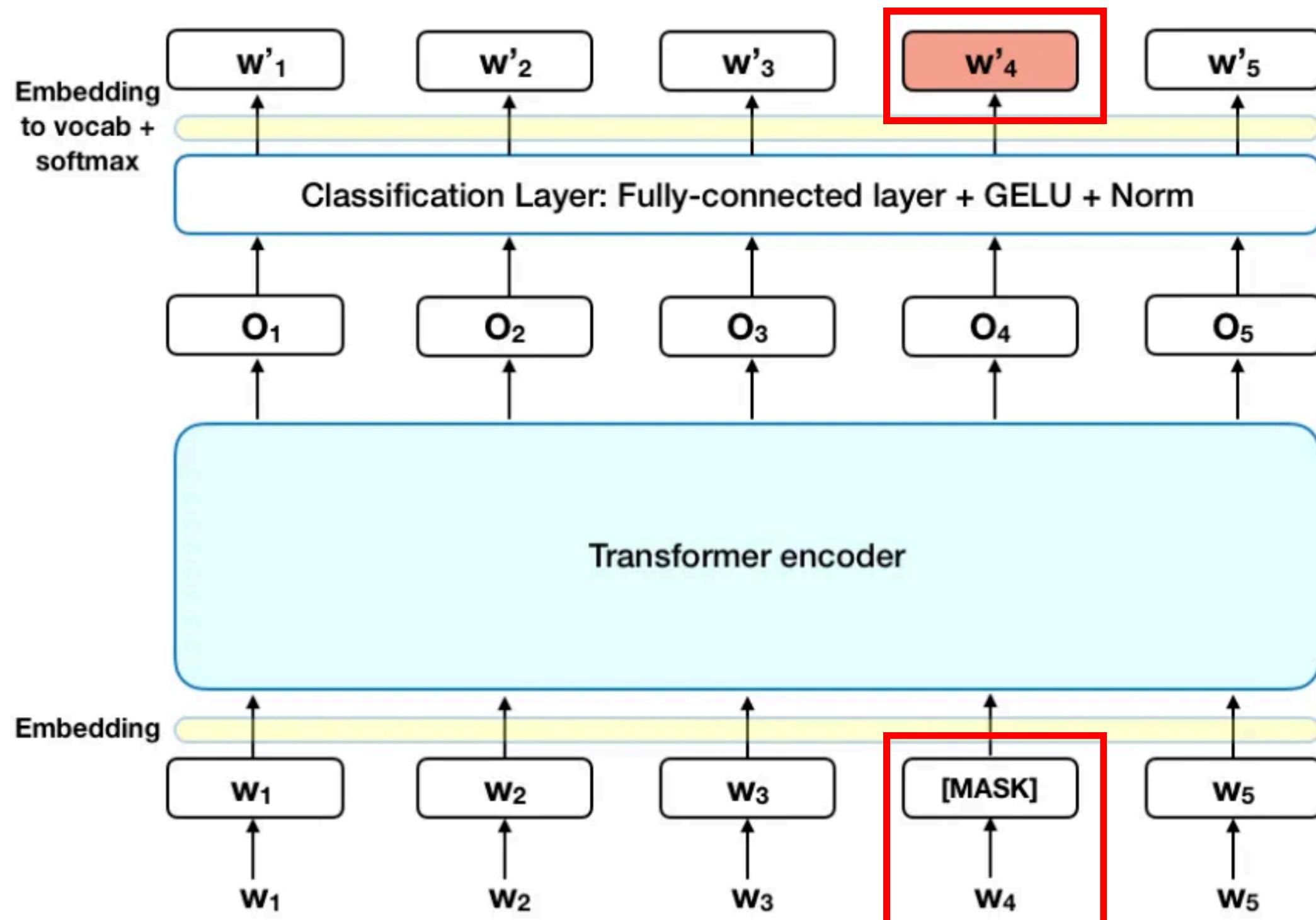
BERT

Pre-training



BERT

Pre-training Tasks



Masked Language Modeling (MLM)

입력 토큰의 15%를 masked 토큰으로 바꿈

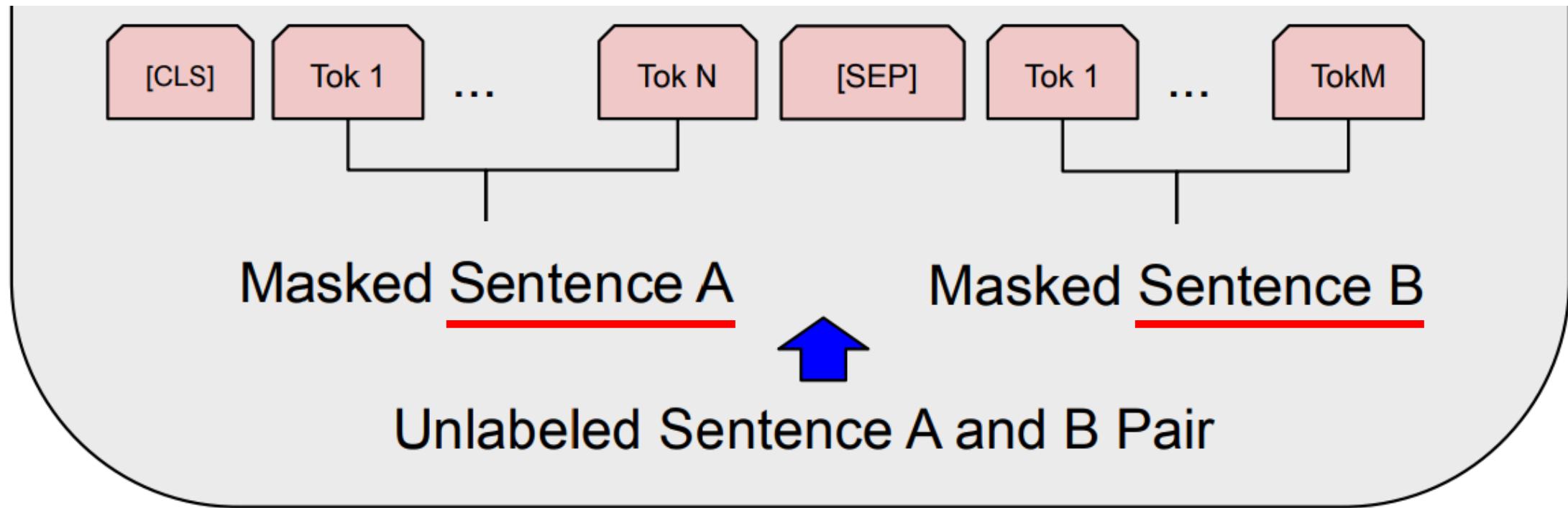
마스크로 선택된 토큰에 대해
80%는 MASK, 10%는 랜덤 토큰, 10% 원토큰

MASK	Masking Rates			Dev Set Results		
	SAME	RND	MNLI	Fine-tune	NER	
			Fine-tune	Feature-based		
80%	10%	10%	84.2	95.4	94.9	
100%	0%	0%	84.3	94.9	94.0	
80%	0%	20%	84.1	95.2	94.6	
80%	20%	0%	84.4	95.2	94.7	
0%	20%	80%	83.7	94.8	94.6	
0%	0%	100%	83.6	94.9	94.6	

Table 8: Ablation over different masking strategies.

BERT

Pre-training Tasks



Next Sentence Prediction (NSP)

두개 이상의 문장의 관계를
알아야 풀 수 있는 task를 위함

	Y
[CLS] S1 [SEP] S2	→ 1
[CLS] S2 [SEP] S3	→ 1
[CLS] S3 [SEP] S6	→ 0
[CLS] S5 [SEP] S10	→ 0

텍스트 쌍(A,B)에 대해
50%는 실제 다음 문장 (**IsNext**),
50%는 임의 문장 (**NotNext**)
[CLS] 표현으로 이진 분류.

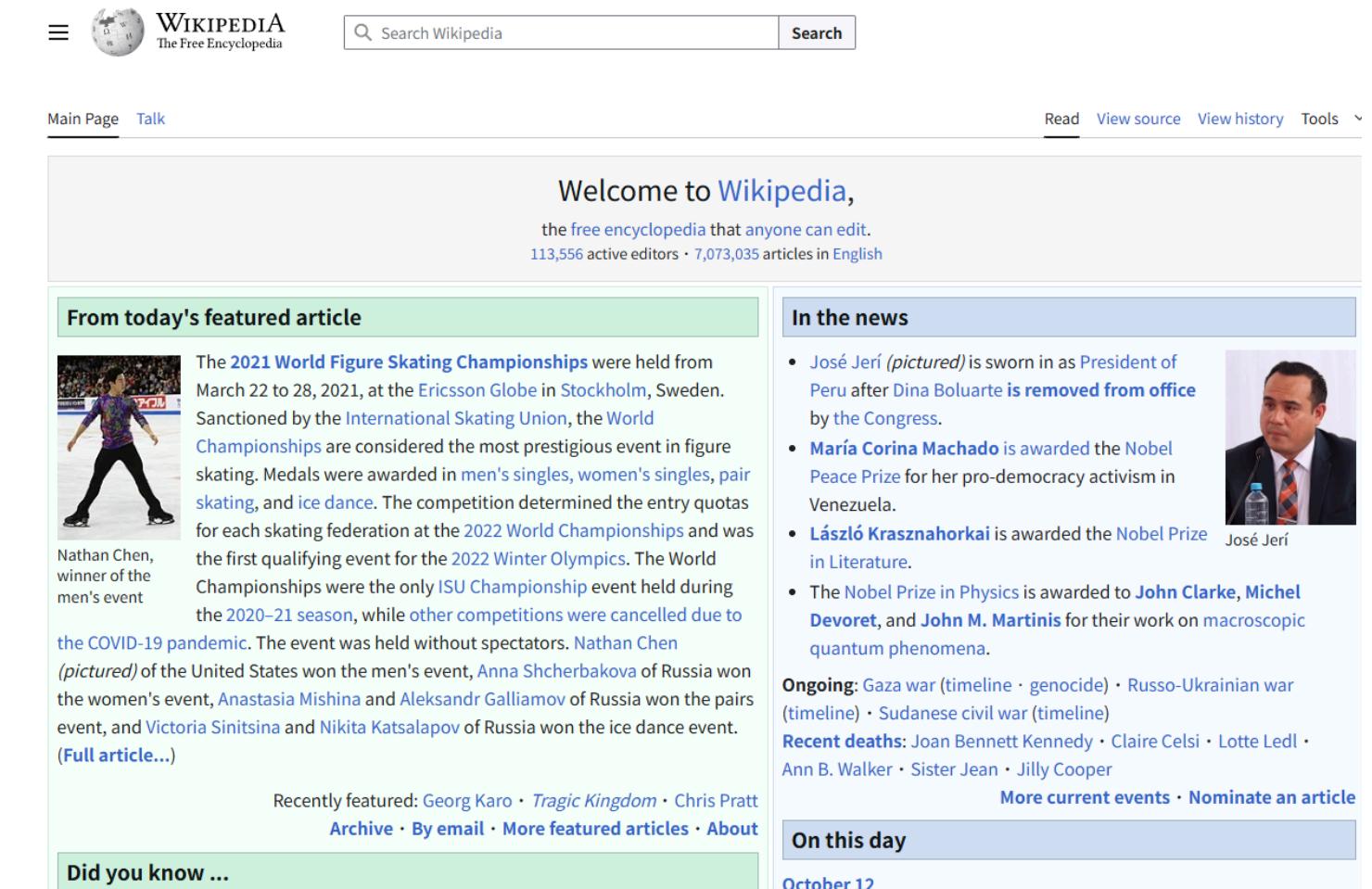
Pre-training Datasets

BookCorpus

Dataset Facts	
Dataset BookCorpus	
Instances Per Dataset 7,185 unique books, 11,038 total	
Motivation	
Original Authors Zhu and Kiros et al. (2015) [39]	
Original Use Case Sentence embedding	
Funding Google, Samsung, NSERC, CIFAR, ONR	
Composition	
Sample or Complete Sample, ≈2% of smashwords.com in 2014	
Missing Data 98 empty files, ≤655 truncated files	
Sensitive Information Author email addresses	
Collection	
Sampling Strategy Free books with ≥20,000 words	
Ethical Review None stated	
Author Consent None	
Cleaning and Labeling	
Cleaning Done None stated, some implicit	
Labeling Done None stated, genres by smashwords.com	
Uses and Distribution	
Notable Uses Language models (e.g. GPT [29], BERT [9])	
Other Uses List available on HuggingFace [12]	
Original Distribution Author website (now defunct) [39]	
Replicate Distribution BookCorpusOpen [13]	
Maintenance and Evolution	
Corrections or Erratum None	
Methods to Extend "Homemade BookCorpus" [21]	
Replicate Maintainers Shawn Presser [12]	
Genres % of BookCorpus*	
Romance 2,881 books	26.1%
Fantasy 1,502 books	13.6%
Vampires 600 books	5.4%
Horror 4.1%	Teen 3.9%
Adventure 3.5%	Literature 3.0%
Historical Fiction 1.6%	
Not a significant source of nonfiction.	
* Percentages based on directories in books_txt_full. Some books cross-listed.	

독립 전자책 배포 웹사이트 스매시워즈에서
웹 스크래핑을 통해 수집한 약 7,000권의
출판 도서 텍스트로 구성된 자료 집합
(8억개 정도의 단어)

영어 Wikipedia



The screenshot shows the English Wikipedia homepage. At the top, there's a search bar with the placeholder "Search Wikipedia" and a "Search" button. Below the search bar, the title "WIKIPEDIA The Free Encyclopedia" is displayed. The main content area features a large banner for the "2021 World Figure Skating Championships". Below the banner, there's a "From today's featured article" section featuring a photo of Nathan Chen and text about his victory at the championships. To the right, there's a "In the news" section with several bullet points about political figures like José Jerí and María Corina Machado, and scientific awards like the Nobel Prize in Physics. At the bottom, there's a "Did you know..." section and a "On this day" section for October 12.

목록, 표, 헤더 등은 무시하고
텍스트 구정만 추출
(25억개 정도의 단어)

Pre-training Hyper-parameter

입력 시퀀스 : 총 길이는 512 토큰을 넘지 않도록 조절

Batch size : 256 시퀀스 (총 128,000 토큰/배치)

총 학습 스텝 : 1,000,000 스텝 (약 33억 단어 코퍼스를 40 에폭(epoch) 학습한 분량)

Optimizer : Adam, Learning rate = $1e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$

(처음 10,000 스텝 동안은 학습률을 서서히 올리고, 그 이후에는 선형적으로 감소)

L2 Weight decay : 0.01

Dropout : 모든 레이어에 0.1 확률로 적용

Activation function : 표준 ReLU 대신 GELU 사용 (OpenAI GPT와 동일)

Loss Function : MLM 손실과 NSP 손실의 평균을 합산하여 사용

Pre-training Procedure

BERT-BASE: 16개의 TPU 칩으로 4일 소요

BERT-LARGE: 64개의 TPU 칩으로 4일 소요

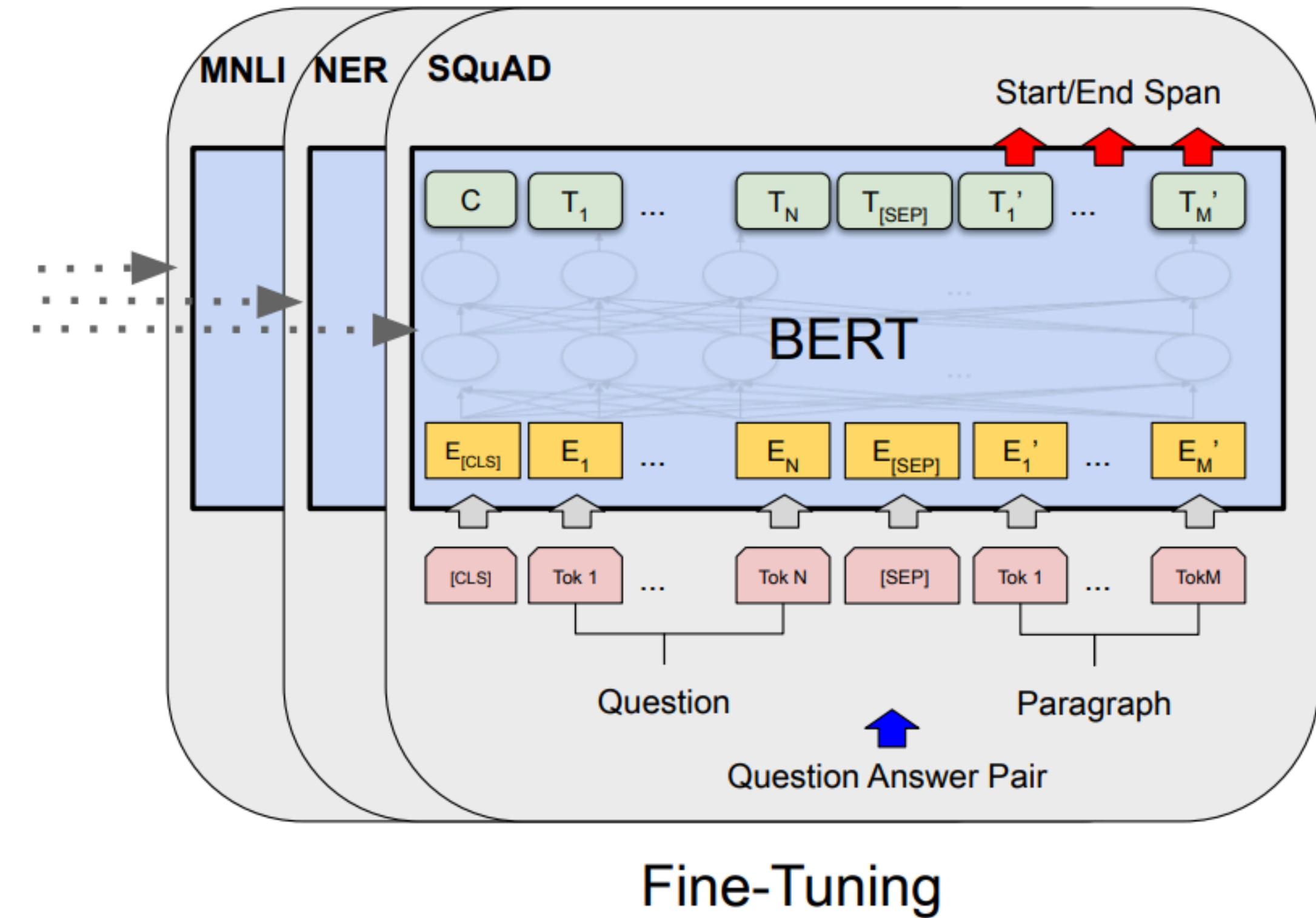
어텐션 메커니즘은 시퀀스 길이가 길어질수록 계산 비용이 제곱으로 증가 → **효율화 필요**

1단계 (90% 진행): 전체 학습 스텝의 90%는 시퀀스 길이를 128로 짧게 하여 빠르게 학습

2단계 (10% 진행): 나머지 10%는 시퀀스 길이를 512로 늘려서 위치 임베딩을 학습

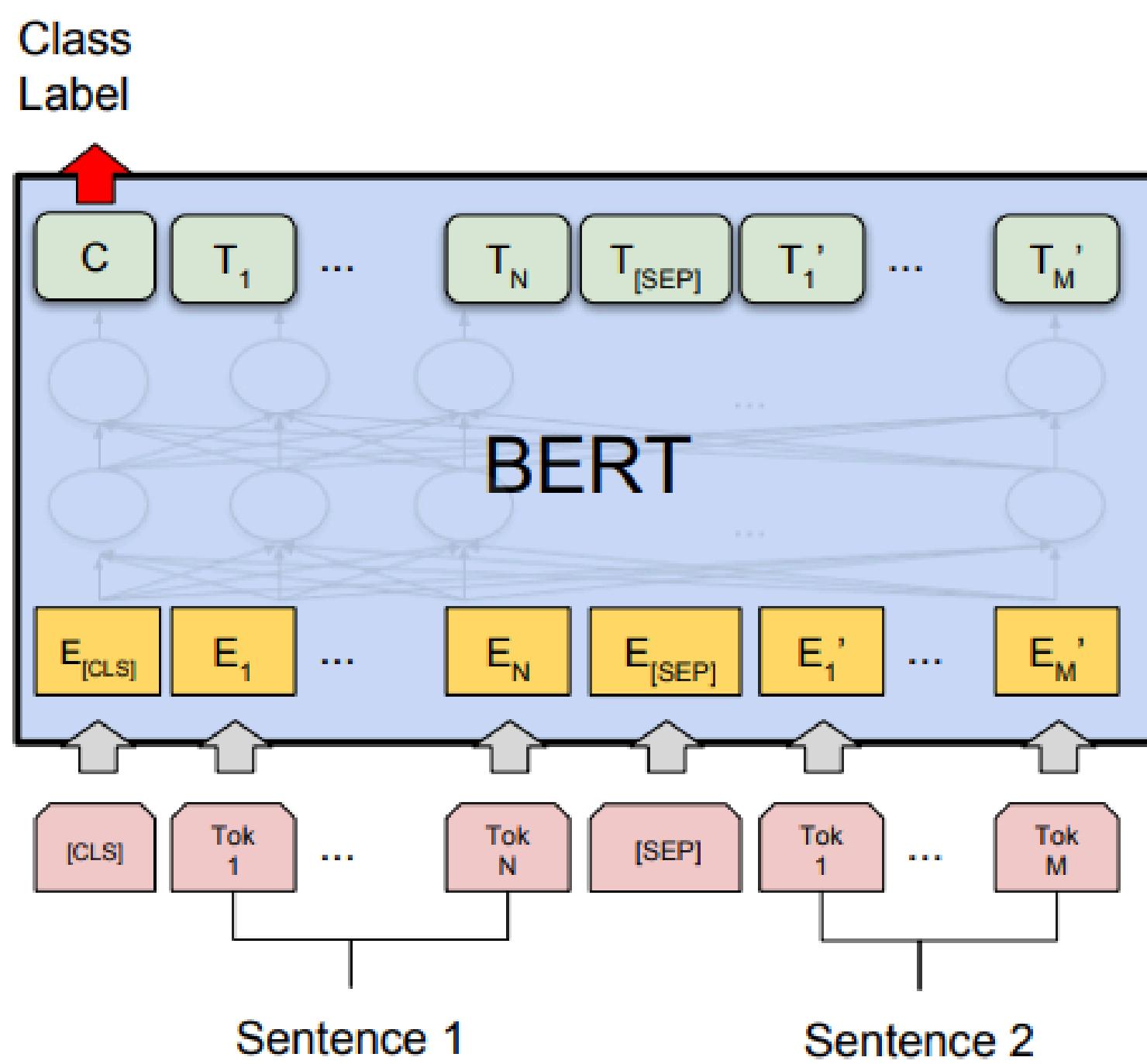
BERT

Fine-tuning



미리 학습된 BERT 모델은 그대로 두고, 그 위에 아주 출력층(output layer) 하나만 추가

Fine-tuning



(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

마지막 레이어를 통과한 [CLS] 토큰의 최종 출력값만 가져와 정답을 예측

MNLI (Multi-Genre Natural Language Inference)

- 두 문장(전제, 가설)이 주어졌을 때, 가설이 전제에 대해 함의(entailment), 모순(contradiction), 중립(neutral) 관계인지 맞추는 3지선다 문제

QQP (Quora Question Pairs)

- Q&A 사이트인 Quora에 올라온 두 개의 질문이 의미적으로 같은 질문인지 아닌지를 맞추는 이진 분류 문제

QNLI (Question Natural Language Inference)

- 질문(Question)과 문단 속 한 문장(Sentence)이 주어졌을 때, 그 문장이 질문에 대한 정답을 포함하는지 아닌지를 맞추는 이진 분류 문제 (스탠포드 QA 데이터셋을 변형)

STS-B (Semantic Textual Similarity Benchmark)

- 뉴스 헤드라인 등에서 추출된 두 문장이 의미적으로 얼마나 유사한지를 1점에서 5점 사이의 점수로 평가하는 문제 (분류가 아닌 회귀(Regression) 문제)

MRPC (Microsoft Research Paraphrase Corpus)

- 온라인 뉴스에서 추출된 두 문장이 의미적으로 같은 내용인지(paraphrase) 아닌지를 판단하는 이진 분류 문제

RTE (Recognizing Textual Entailment)

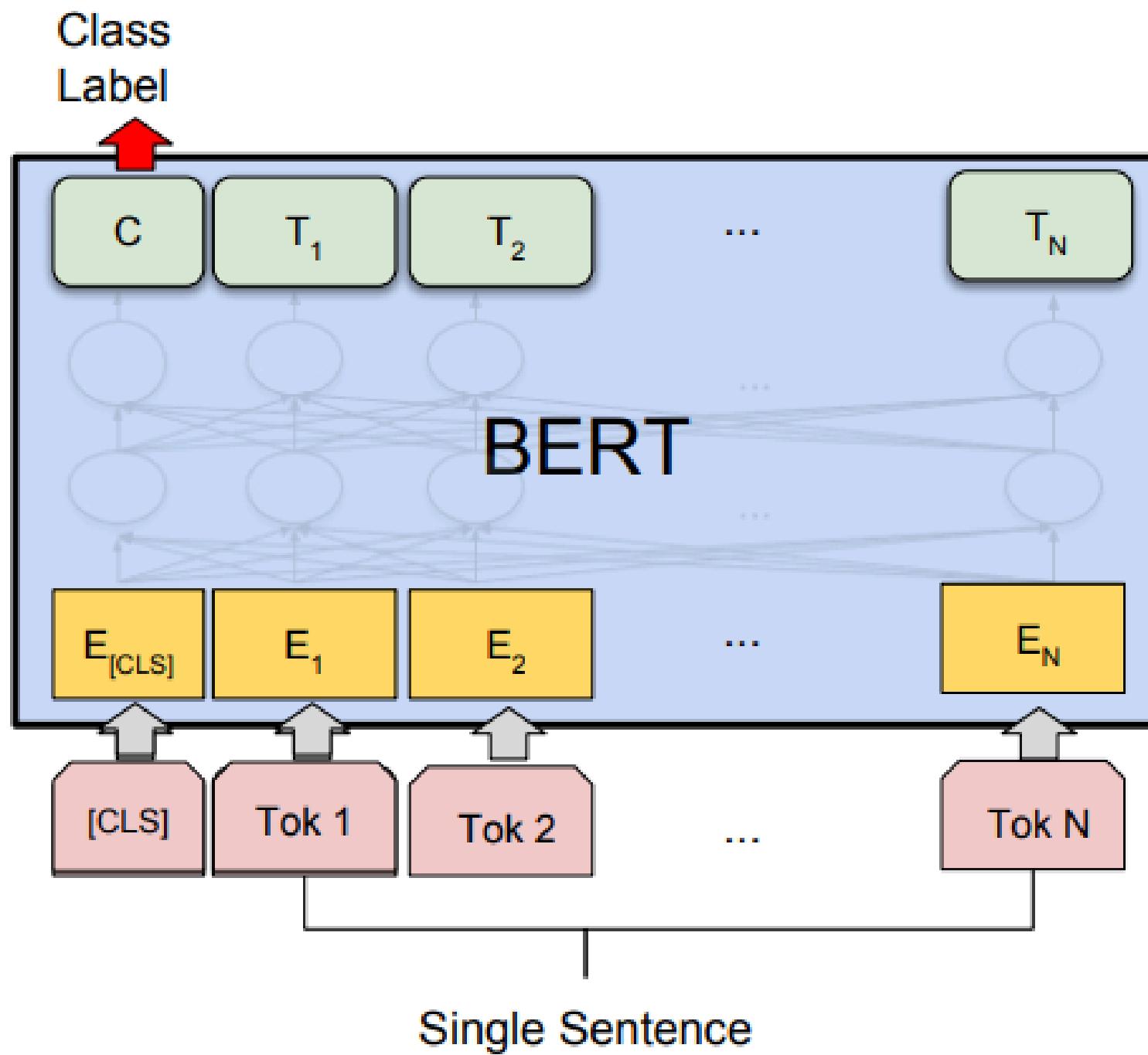
- MNLI와 유사한 함의 관계 추론 문제이지만, 학습 데이터의 양이 훨씬 적어서 더 어려운 과제

SWAG (Situations With Adversarial Generations)

- 인간은 잘풀지만 AI는 잘 풀지 못하는 Adversarial Filtering으로 오답 보기들을 만들어낸 과제

BERT

Fine-tuning



마지막 레이어를 통과한 [CLS] 토큰의 최종 출력값만 가져와 정답을 예측

SST-2 (Stanford Sentiment Treebank)

- 영화 리뷰 문장이 긍정적인지 부정적인지를 판단하는 이진 감성 분류 문제

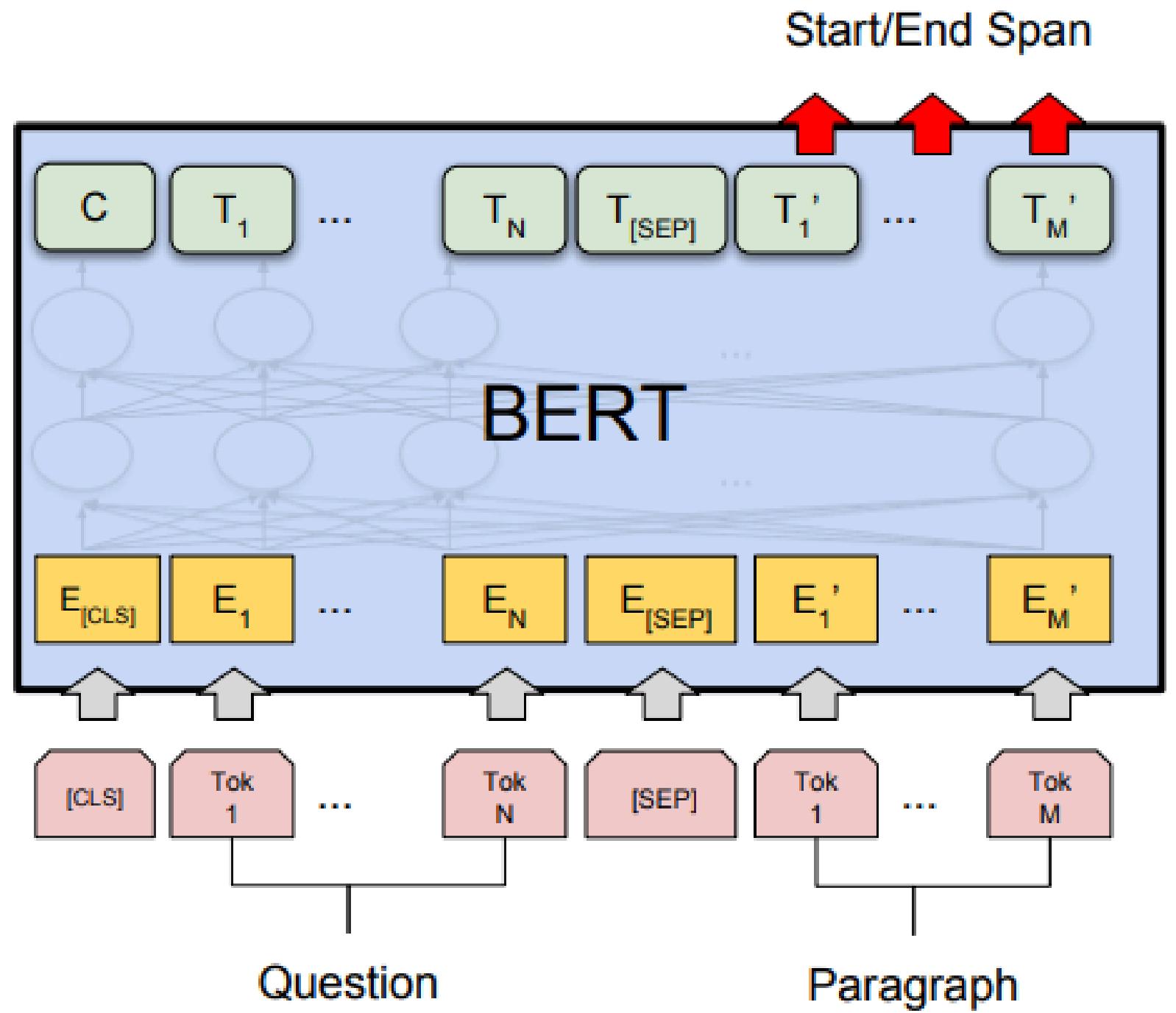
CoLA (Corpus of Linguistic Acceptability)

- 주어진 영어 문장이 문법적으로 올바른 문장인지 아닌지를 판단하는 이진 분류 문제

(b) Single Sentence Classification Tasks:
SST-2, CoLA

BERT

Fine-tuning



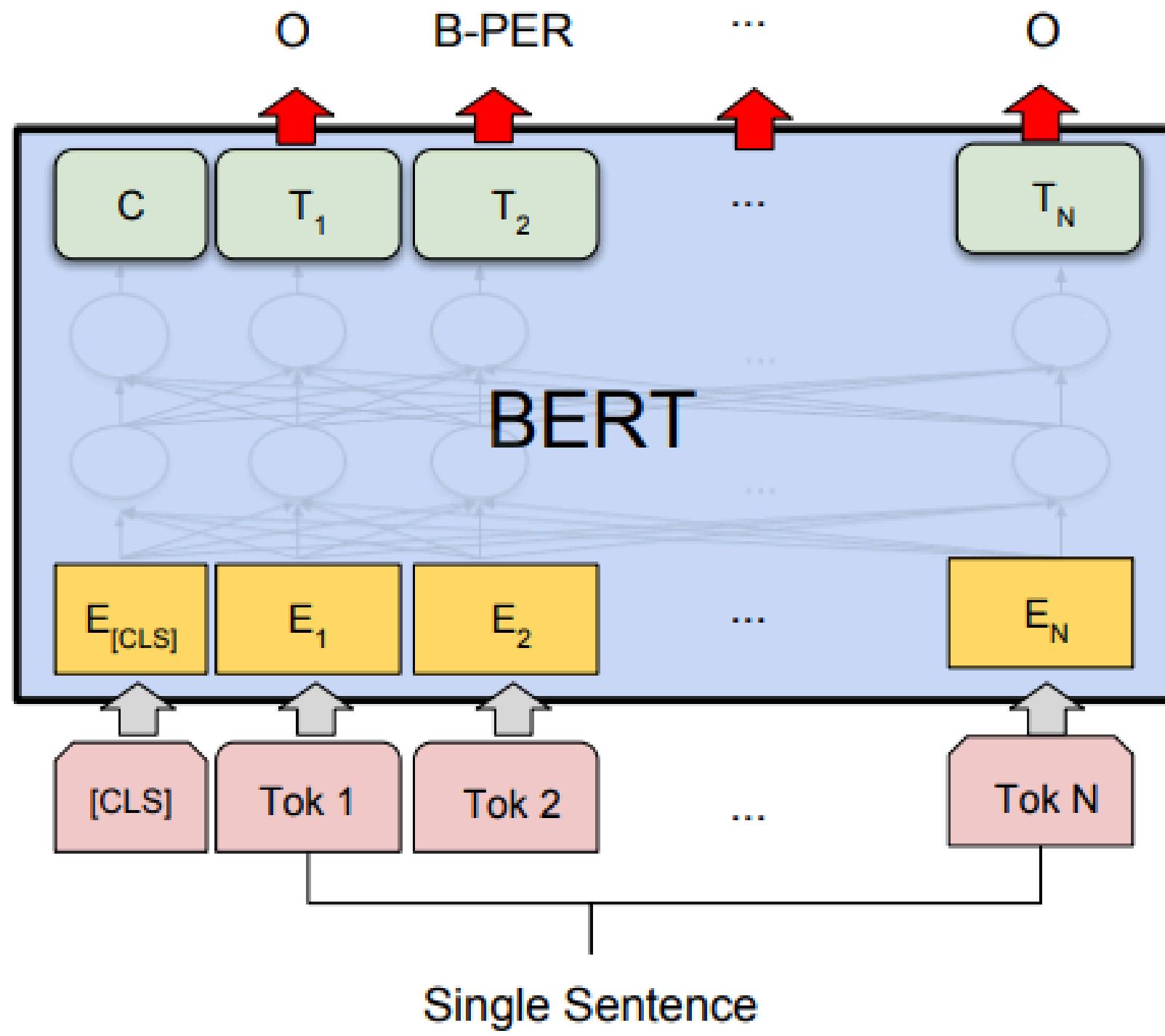
[CLS] 토큰을 사용하지 않음
질문에 대한 정답이 본문 어디에 있는지 시작 토큰과 끝 토큰을
찾아내는 문제

SQuAD(The Stanford Question Answering Dataset)
영문 위키피디아에서 페이지랭크 기준 상위 문서중 무작위로 추출 한다음 사람이 직접 작업해
고품질의 질문과 답변으로 구성한 대규모 데이터셋

(c) Question Answering Tasks:
SQuAD v1.1

BERT

Fine-tuning



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

[CLS] 토큰을 사용하지 않음
대신 마지막 레이어를 통과한 모든 토큰 각각의 최종 출력값을 사용

CoNLL-2003 NER(Conference on Computational Natural Language Learning, Named Entity Recognition)

- 문장에서 인물, 장소, 기관 등과 같이 이름(고유명사)을 가진 대상을 찾아내고 그 종류를 분류하는 자연어 처리의 핵심 과제 중 하나
- 로이터(Reuters) 뉴스 기사에서 수집된 텍스트를 사용
- 영어와 독일어 두 가지 언어의 데이터셋을 제공

개체명 유형: 총 4가지의 주요 개체명 유형을 분류합니다.

- PER: 인물 (Person) - 예: Harry Potter, Einstein
- LOC: 장소 (Location) - 예: Seoul, Germany
- ORG: 기관 (Organization) - 예: Google, United Nations
- MISC: 기타 (Miscellaneous) - 위의 3가지에 속하지 않는 고유명사. 예: Korean, Christmas

데이터 형식: IOB2 포맷을 사용해 각 단어(토큰)에 태그를 붙입니다.

- B-: 개체명이 시작되는 단어 (Begin)
- I-: 개체명의 시작이 아닌 중간 부분 단어 (Inside)
- O: 개체명이 아닌 단어 (Outside)

BERT

Experiment

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Tasks	Dev Set					Hyperparams		Dev Set Accuracy				
	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)	#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
						3	768	12	5.84	77.9	79.8	88.4
BERT _{BASE}	84.4	88.4	86.7	92.7	88.5	6	768	3	5.24	80.6	82.2	90.7
No NSP	83.9	84.9	86.5	92.6	87.9	6	768	12	4.68	81.9	84.8	91.3
LTR & No NSP	82.1	84.3	77.5	92.1	77.8	12	768	12	3.99	84.4	86.7	92.9
+ BiLSTM	82.1	84.1	75.7	91.6	84.9	12	1024	16	3.54	85.7	86.9	93.3
						24	1024	16	3.23	86.6	87.8	93.7

NSP의 유무에 따른 성능 차이

모델 사이즈에 따른 성능 차이

GPT-1

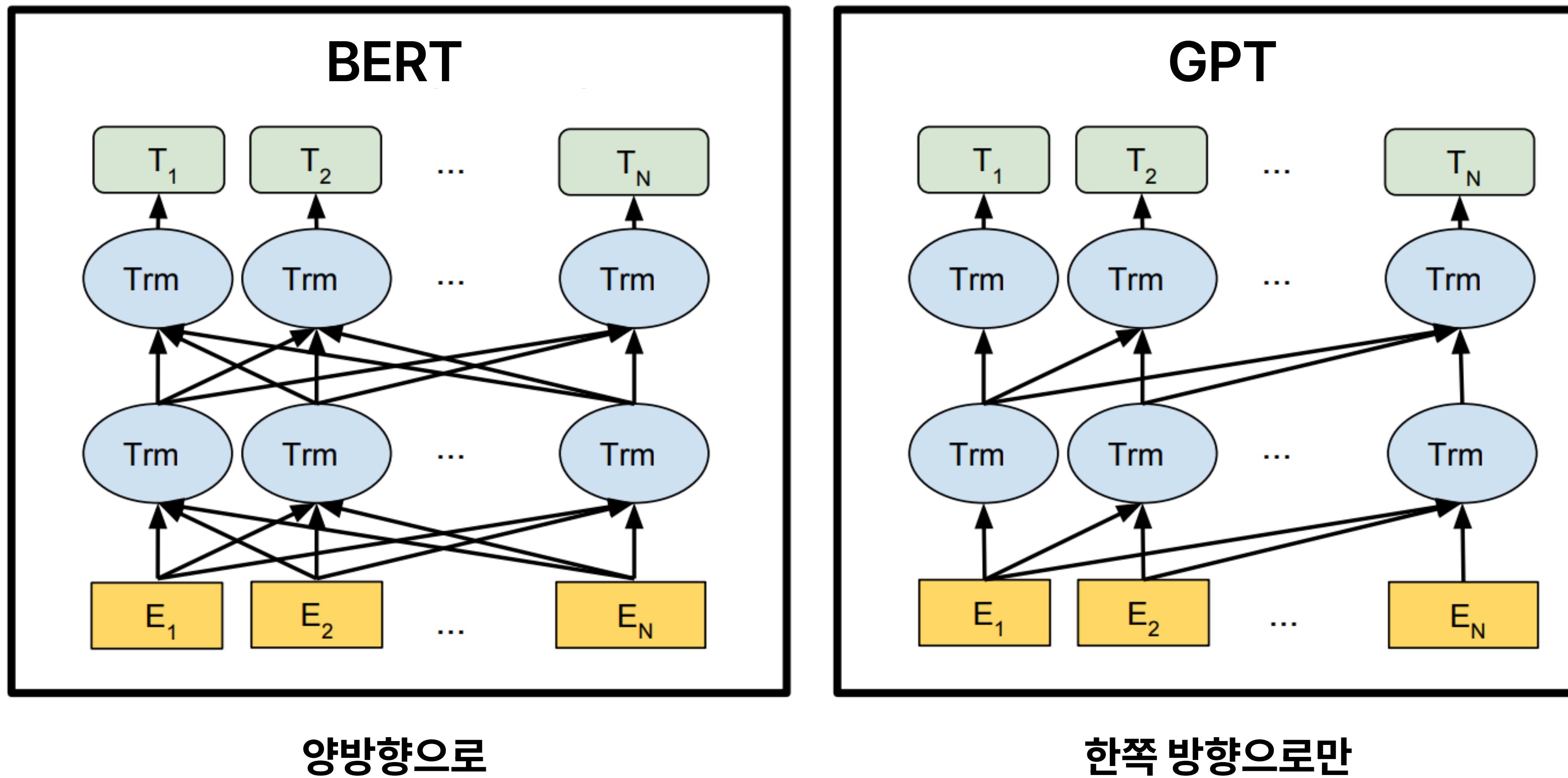
Decoder

Generative
pre-trained
transformer

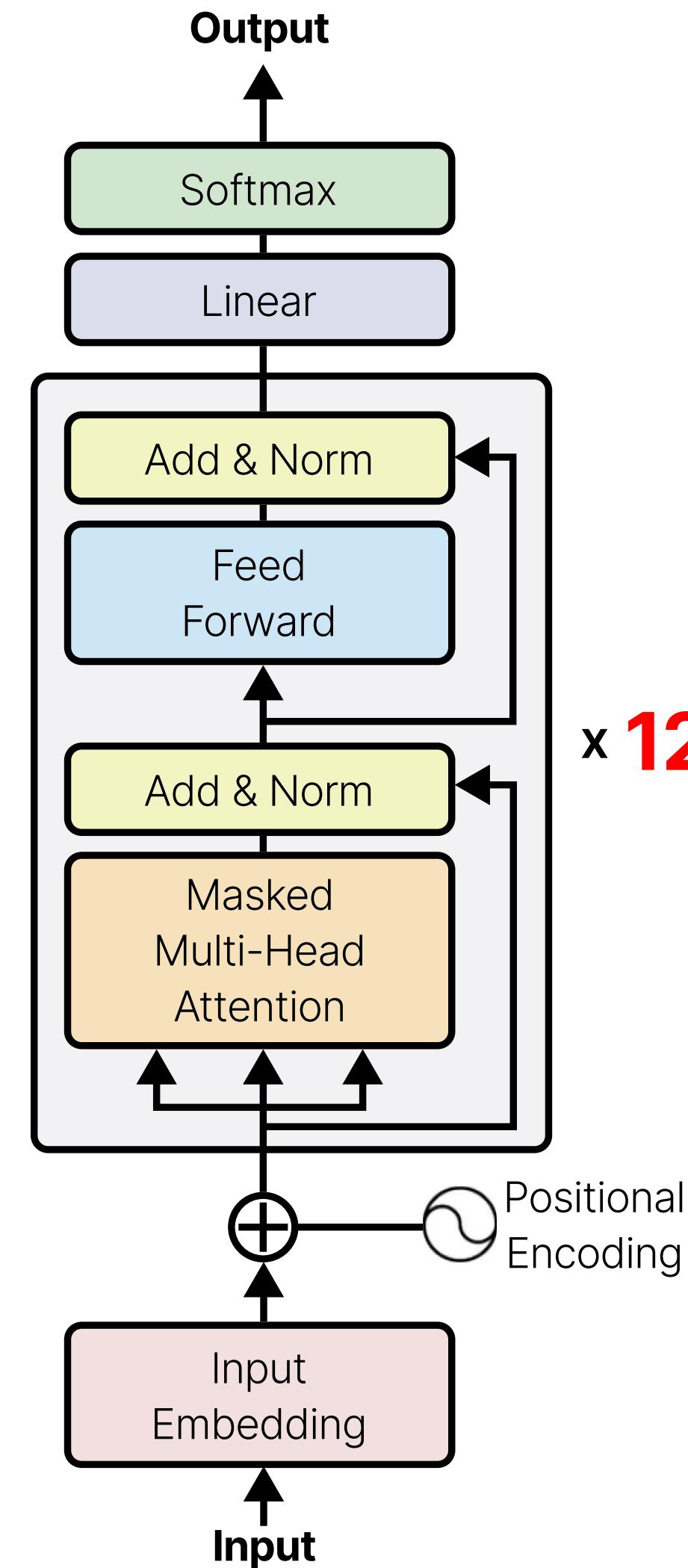


GPT-1

BERT vs. GPT



Architecture



Transformer 구조에서
Encoder와 연결되는 중간 Multi-Head Attention 부분이
제거된 블록이 여러개 연결된 구조

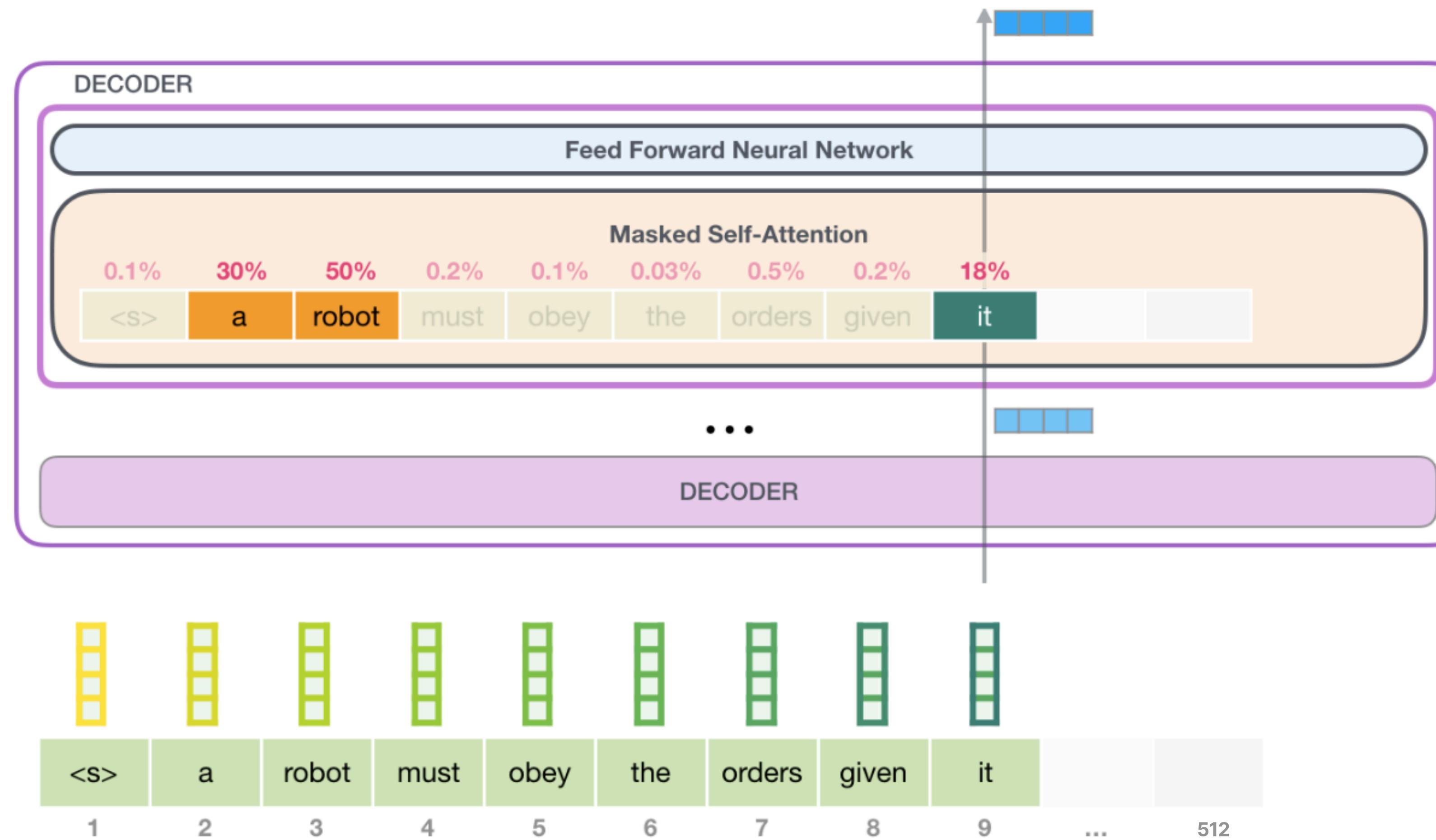
GPT-1

Token embedding & Positional encoding



GPT-1

Self-Attention



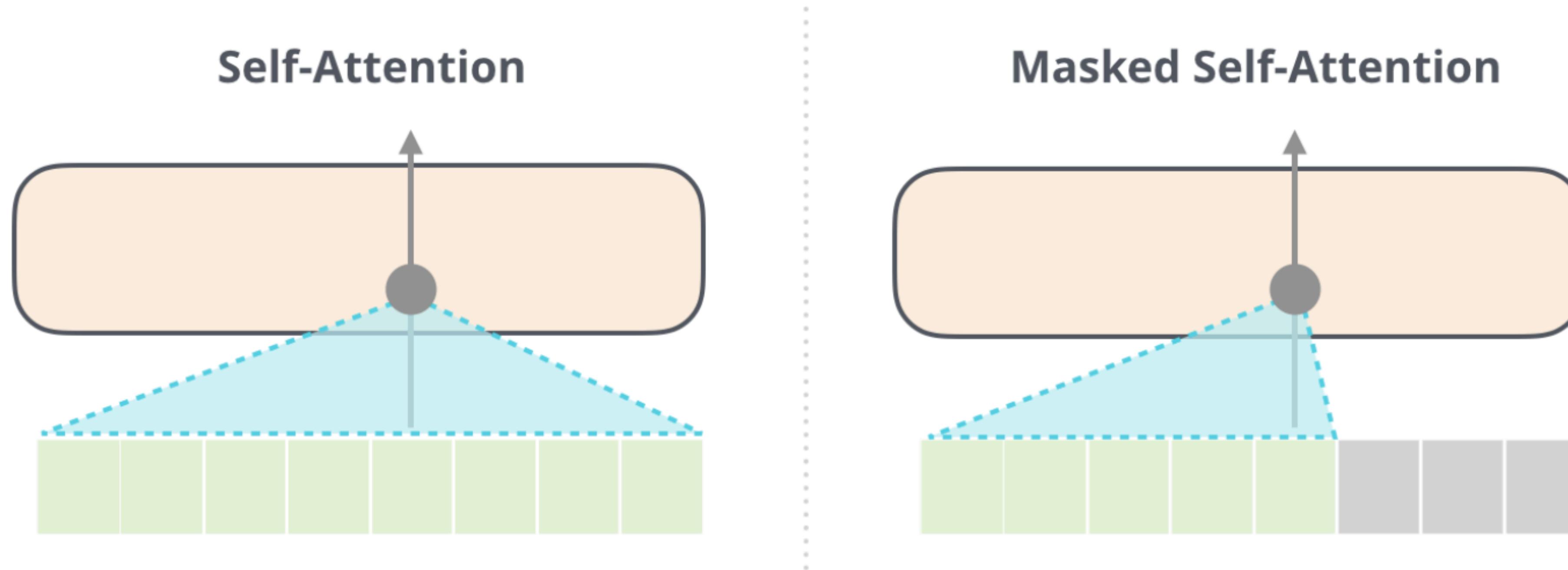
GPT-1

Pre-training

Dataset : BooksCorpus

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta)$$

파라미터를 가진 신경망



Fine-tuning

$$P(y|x^1, \dots, x^m) = \text{softmax}(h_l^m W_y)$$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \dots, x^m)$$

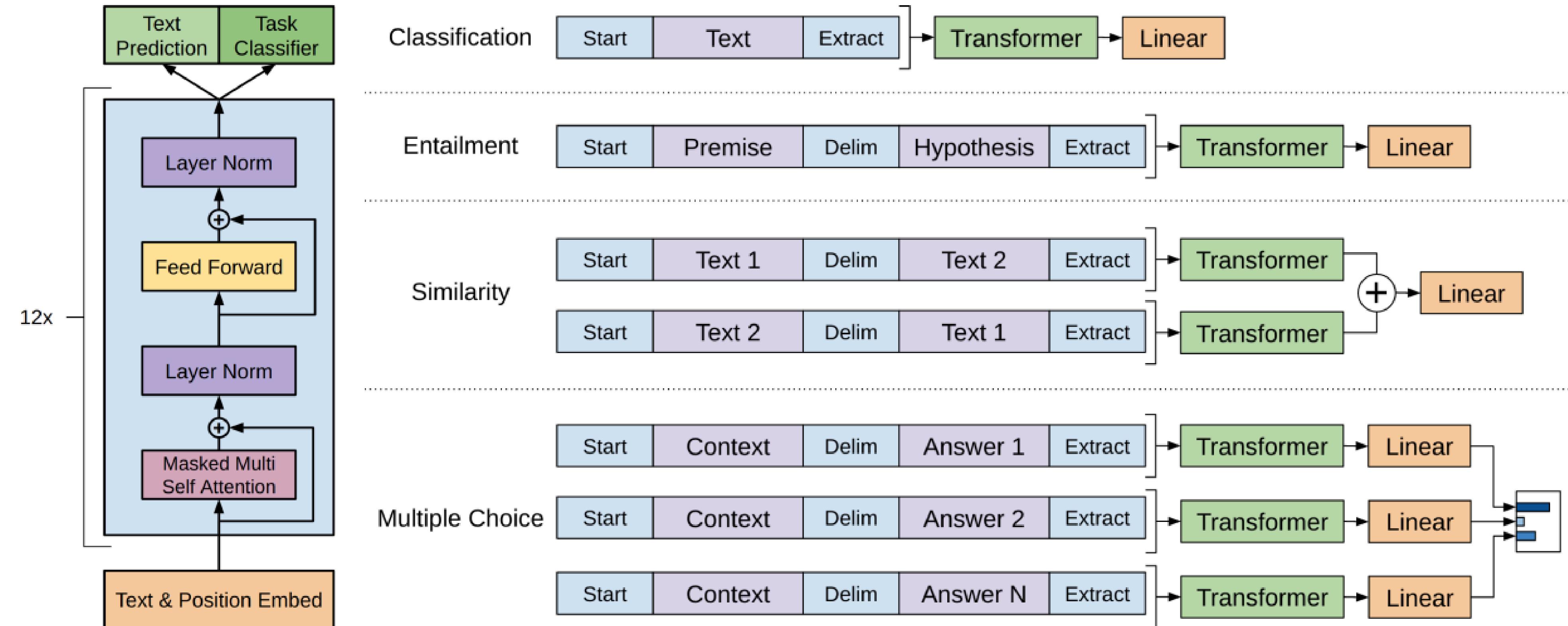
$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

미세 조정을 할 때 원래의 언어 모델링(다음 단어 예측)을 학습에 함께 사용

- 모델이 특정 과제의 데이터에만 과적합되는 것을 막아줌
- 모델이 새로운 과제를 더 빨리 학습

GPT-1

Task에 맞는 input 형태



Experiment

Natural Language Inference

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>		
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	61.7
Finetuned Transformer LM (ours)	82.1	81.4	89.9	88.3	88.1	56.0

Q & A

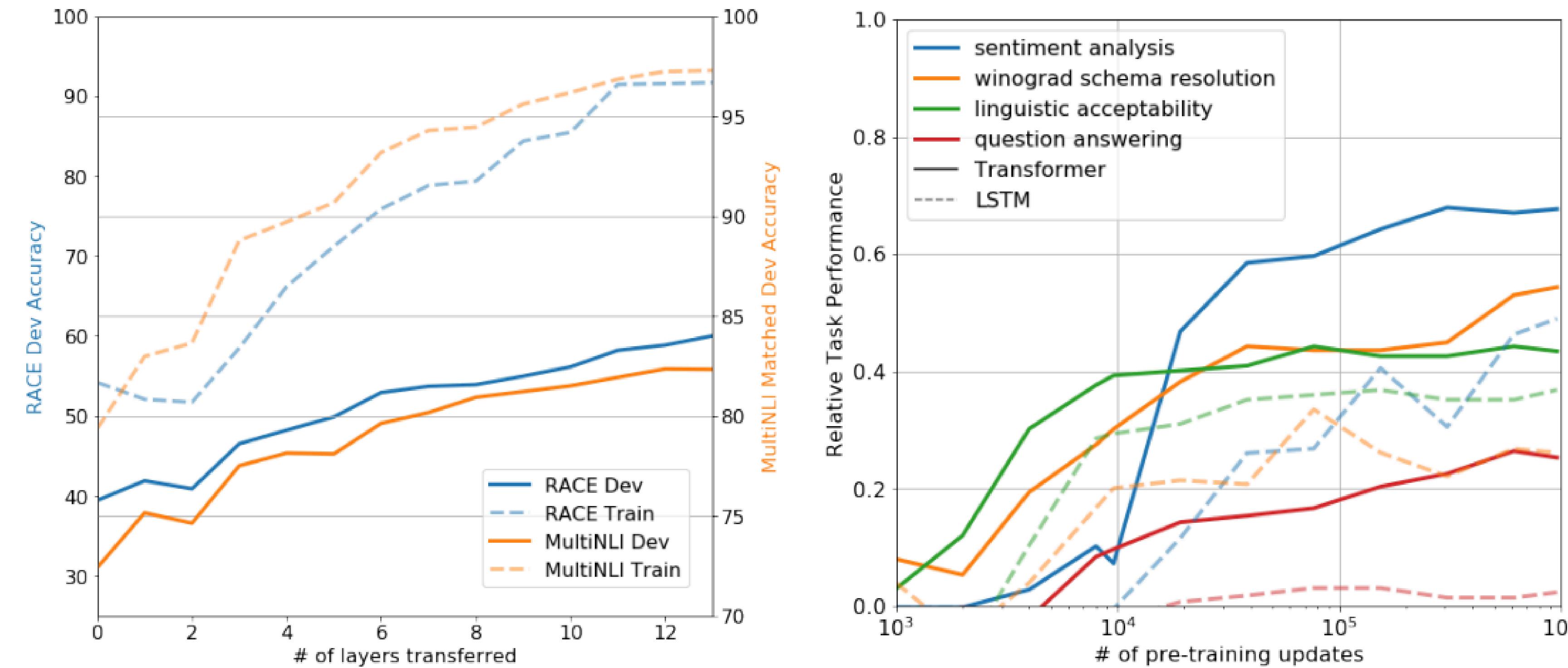
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	<u>77.6</u>	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	<u>60.2</u>	<u>50.3</u>	<u>53.3</u>
Finetuned Transformer LM (ours)	86.5	62.9	57.4	59.0

Experiment

Classification & Semantic Similarity

Method	Classification		Semantic Similarity		GLUE	
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	93.2	-	-	-	-
TF-KLD [23]	-	-	86.0	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	<u>81.0</u>	-	-
Single-task BiLSTM + ELMo + Attn [64]	<u>35.0</u>	90.2	80.2	55.5	<u>66.1</u>	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	<u>68.9</u>
Finetuned Transformer LM (ours)	45.4	91.3	82.3	82.0	70.3	72.8

Experiment



decoding block을
몇개나 쌓아야하는지

비지도 pre-training 후에
zero shot한거랑 fine tuning한거랑 비교

Experiment

Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	70.3	81.8	88.1	56.0
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	75.0	47.9	92.0	84.9	83.2	69.8	81.1	86.9	54.4
LSTM w/ aux LM	69.1	30.3	90.5	83.2	71.8	68.1	73.7	81.1	54.6

보조 목표(Auxiliary Objective) 제거 실험

데이터셋의 크기가 클수록 보조 목표가 일반화 성능을 높이는 데 도움이 되지만,
작은 데이터셋에서는 큰 효과가 없을 수도 있음

Transformer를 LSTM으로 교체

Transformer 구조가 우수함

사전 훈련(Pre-training) 생략

사전 훈련이 모델 성능에 가장 중요한 역할을 해줌

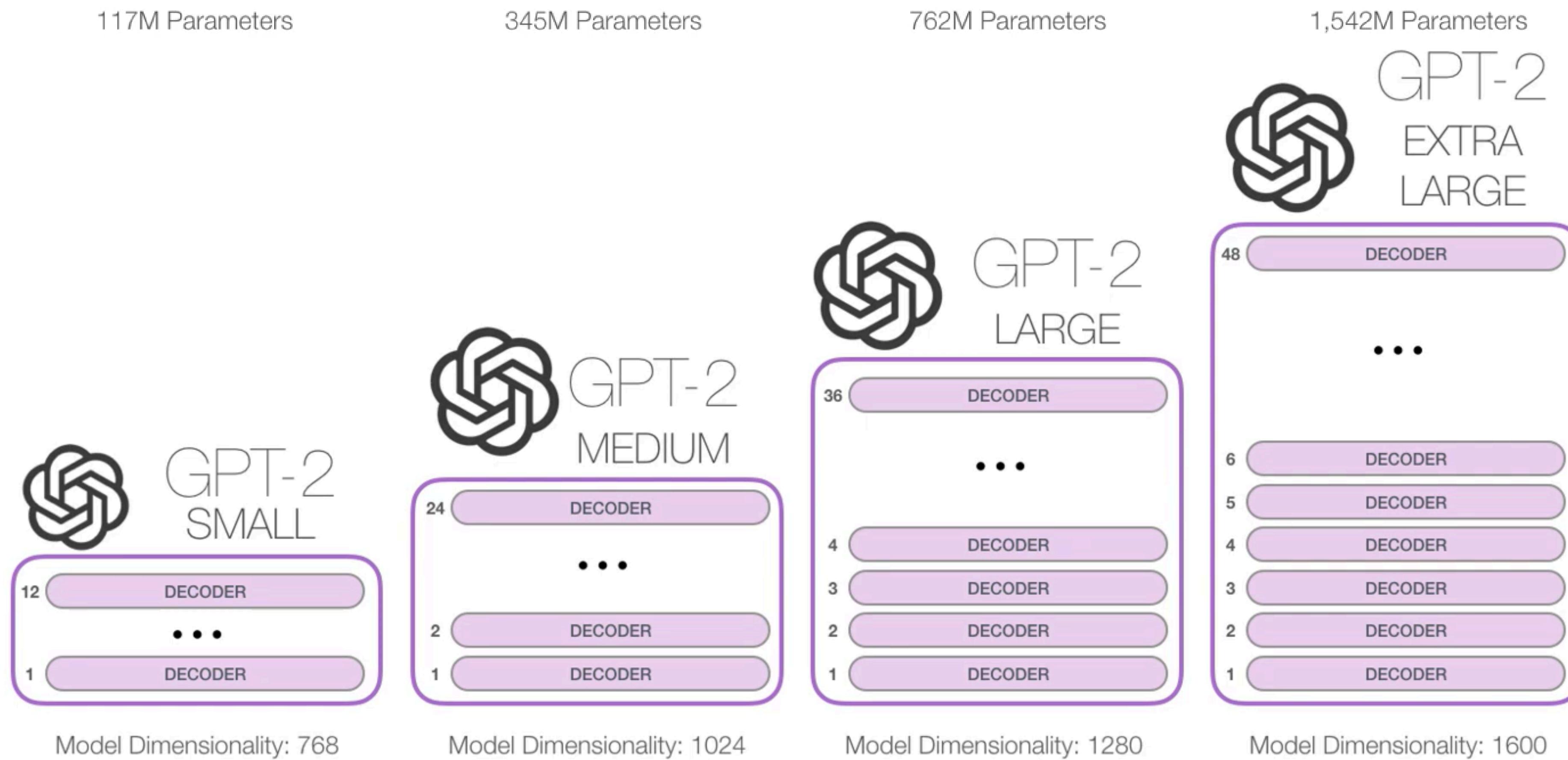
GPT-2

Generative
pre-trained
transformer

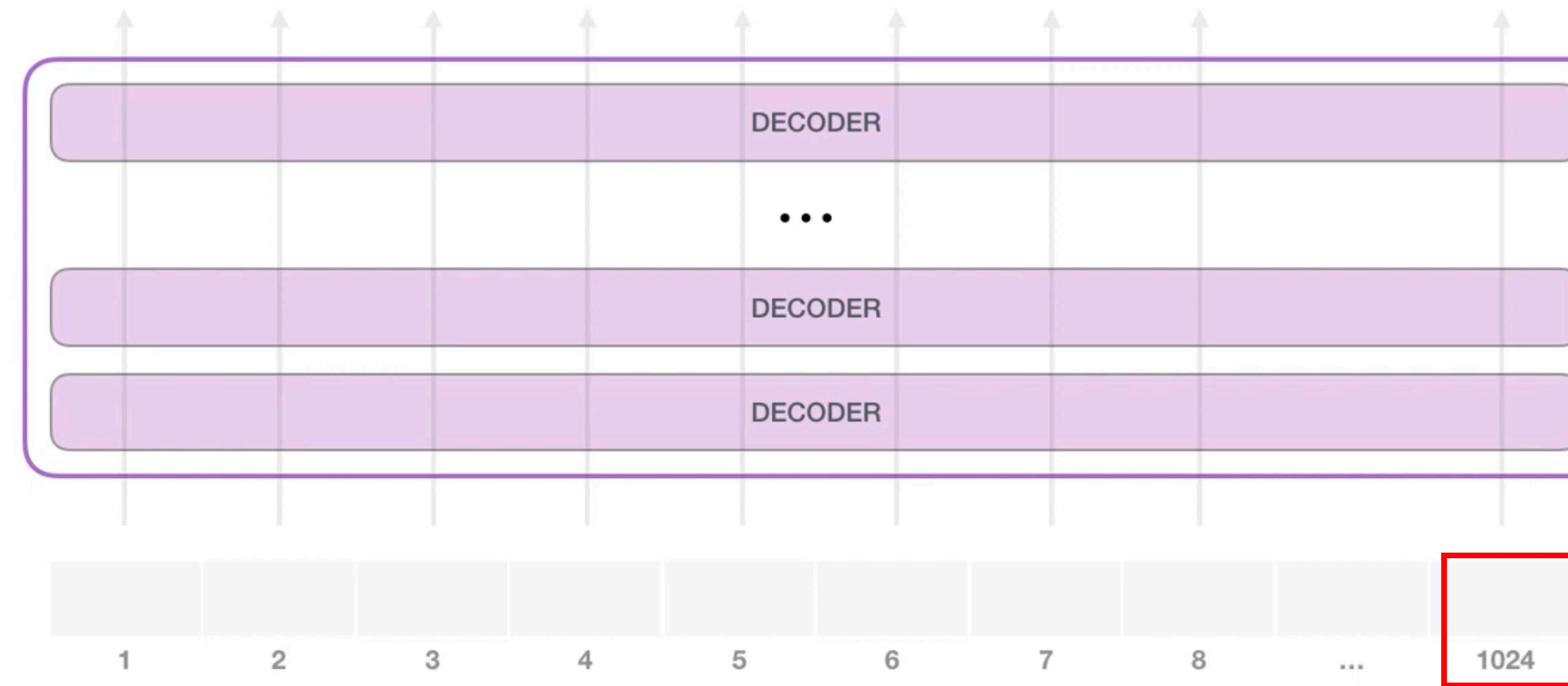


GPT-2

Architecture

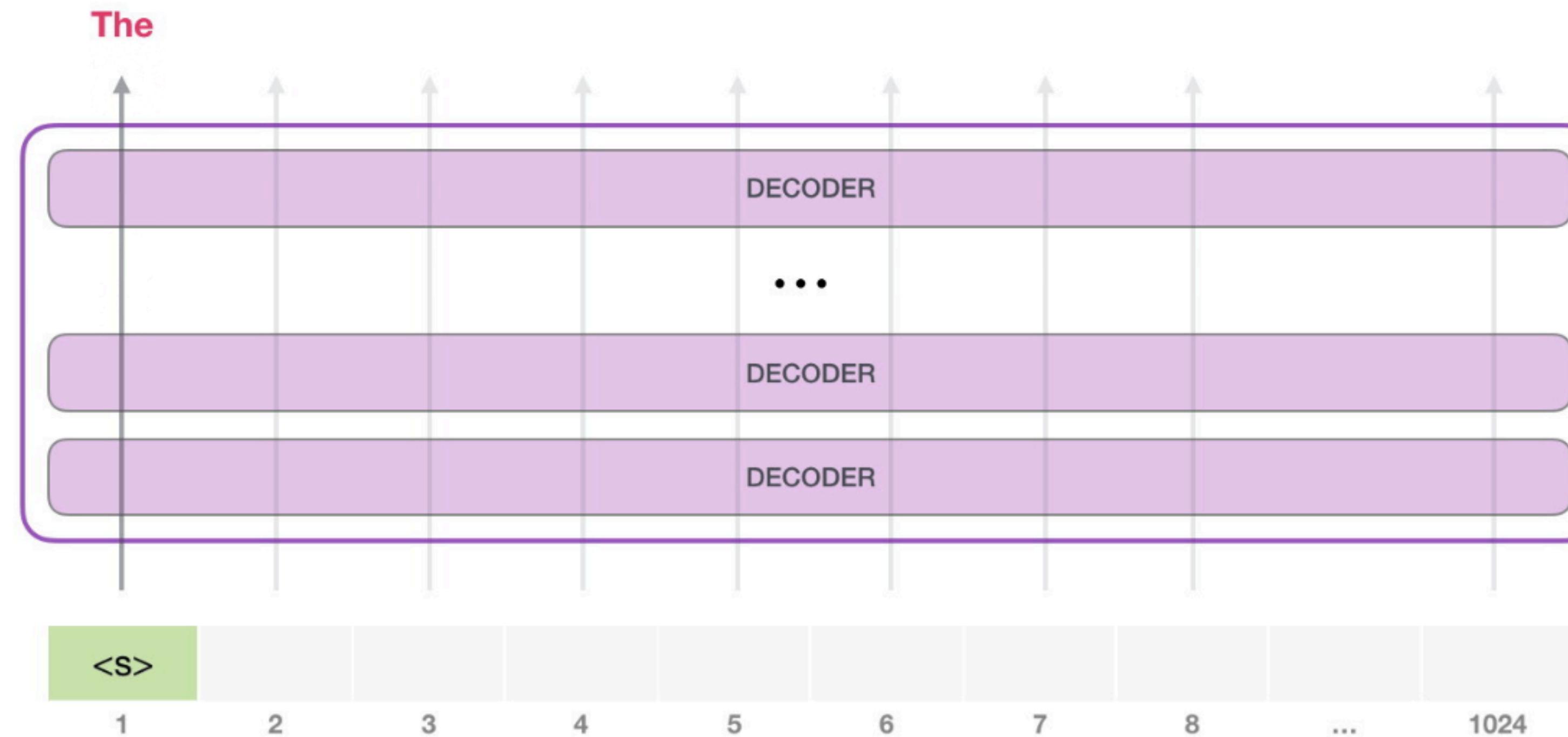


Architecture



1024개의 토큰을 처리할 수 있음

Architecture

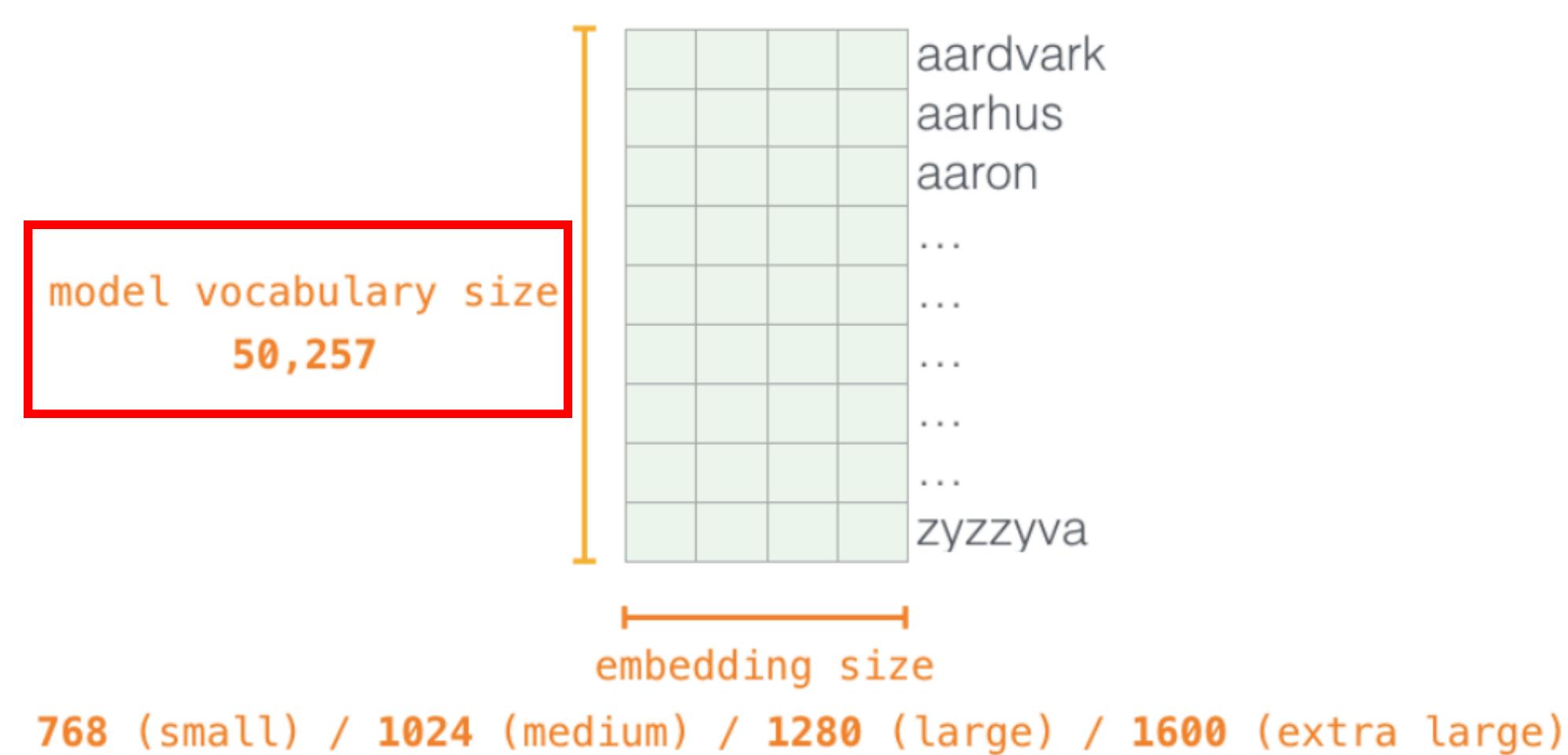


단어를 생성할 때,
GPT-2부터는 top-k 파라미터를 사용해
가장 확률이 높은애를 선택 하는 범위를 설정 가능

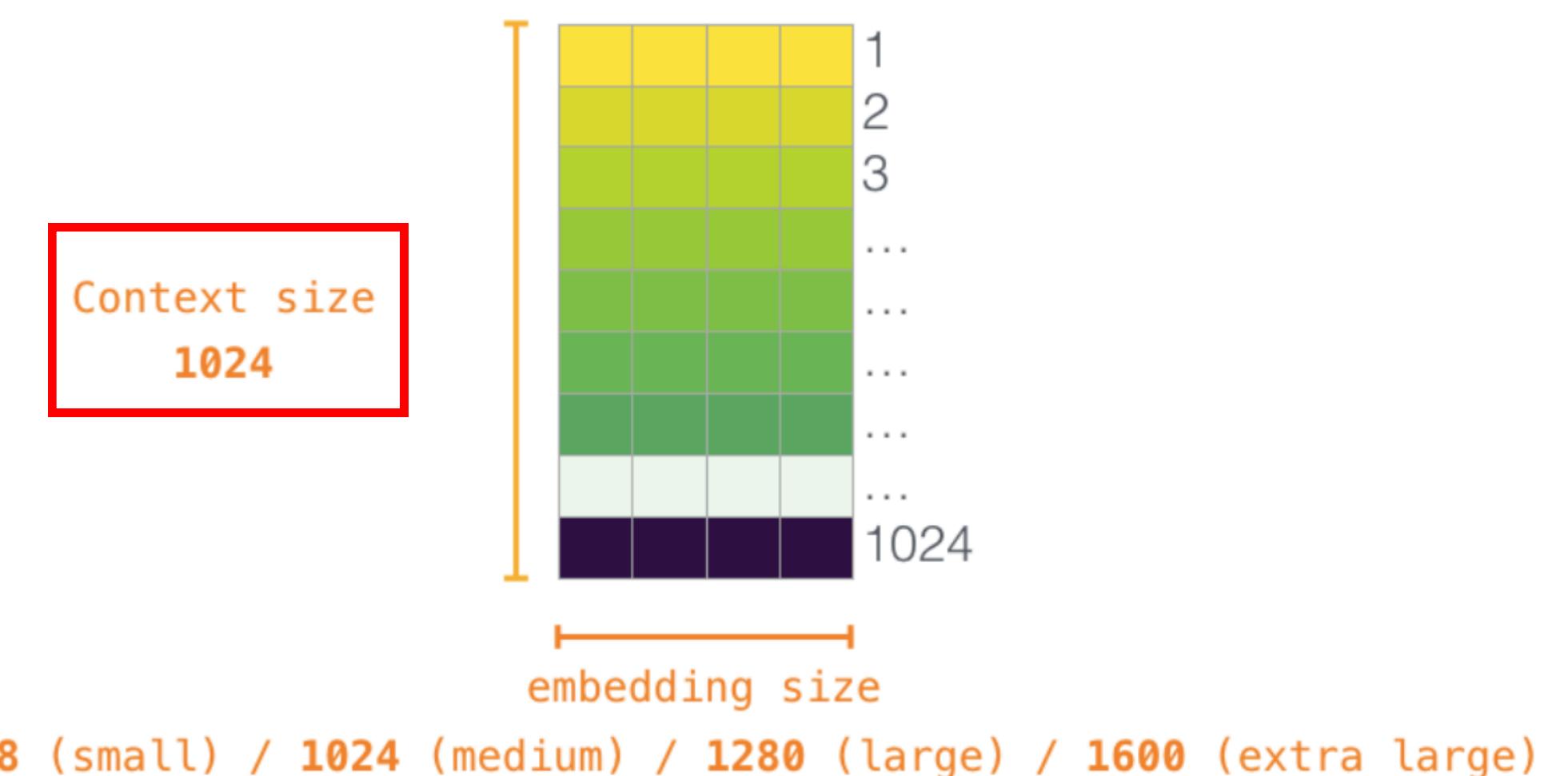
GPT-2

Token embedding & Positional encoding

Token Embeddings (wte)



Positional Encodings (wpe)



Dataset

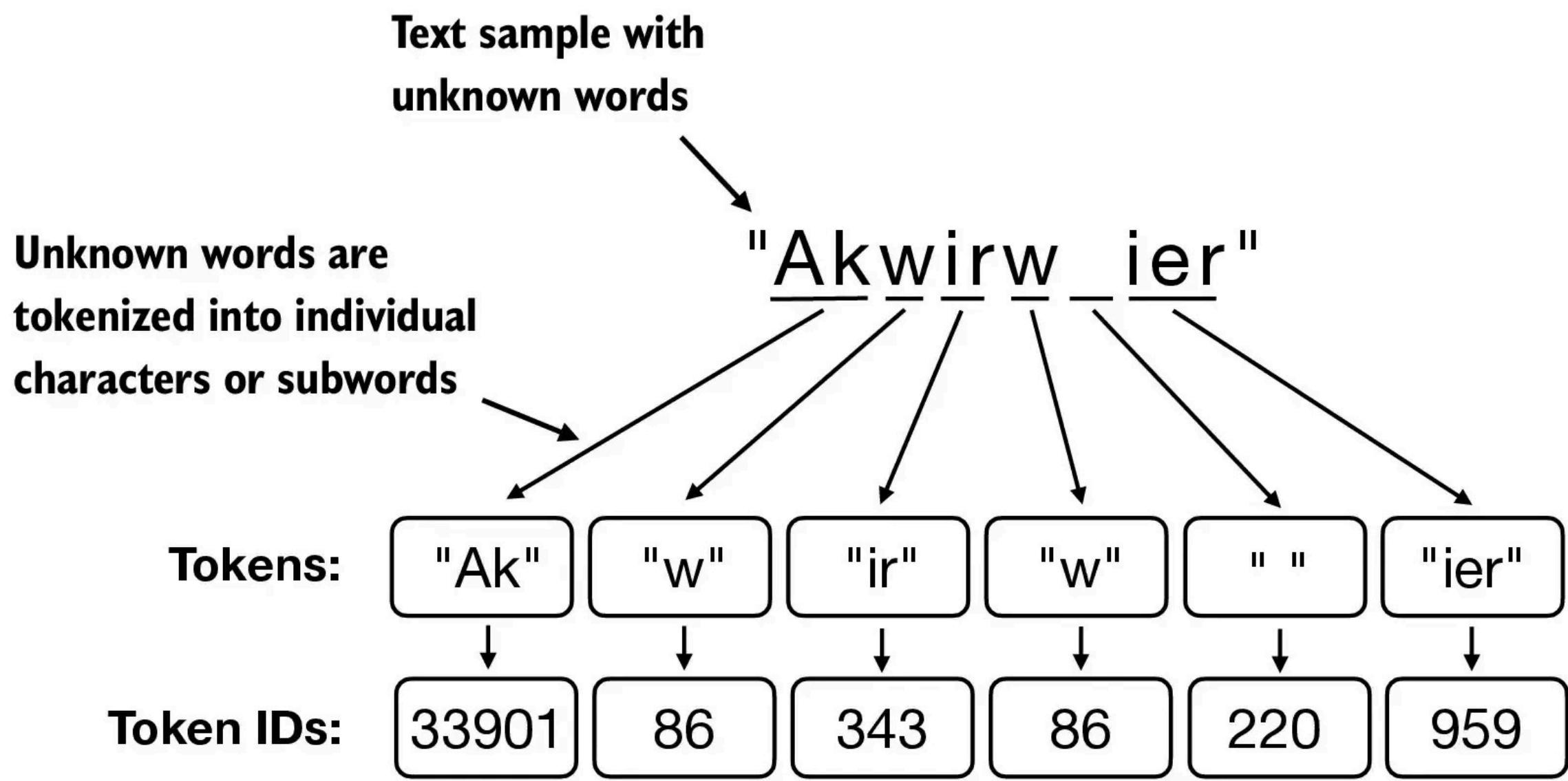
WebText

이전의 언어 모델들은 뉴스, 위키피디아, 소설책 등 특정 분야의 데이터로만 훈련되는 경우가 많았음
다양한 분야와 문맥의 데이터를 최대한 많이 모아, 모델이 여러 과업을 자연스럽게 학습하게 시도함

다양한 도메인의 고품질 웹페이지를 수집

전 웹 스크랩은 품질 문제가 크므로, Reddit에서 [추천 ≥ 3]을 받은 링크만 크롤링
HTML에서 본문만 추출하고, 중복 제거와 정제 후 약 800만 문서, 40GB 텍스트 확보
Wikipedia는 제외해 평가 데이터와의 중복을 줄임

Byte Pair Encoding

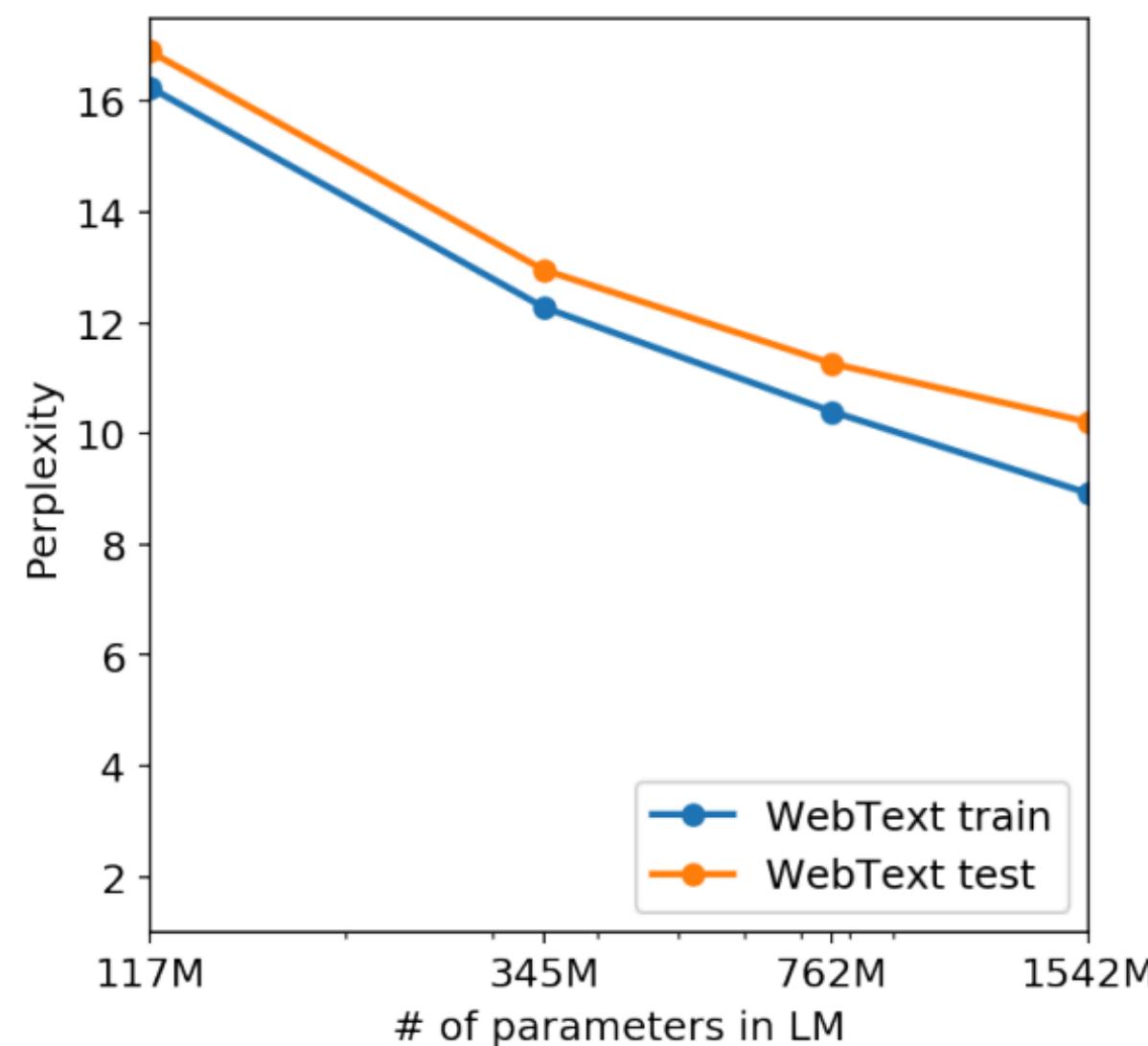


Byte Pair Encoding

자주 같이 나오는 글자들을 합쳐서
하나의 '토큰'으로 만드는
단어 분리 압축 알고리즘

처음 보는 단어(신조어, 오타 등)에
대처하기 위한 핵심 기술

Experiment



파라미터가 많아질수록 perplexity가 떨어짐

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	21.8
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50	75.20
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37	55.72
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05	44.575
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48	42.16

zero-shot 만으로도 SOTA 급 성능

Experiment

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Training Dataset에 없는 질문에 잘 대답함

GPT-3

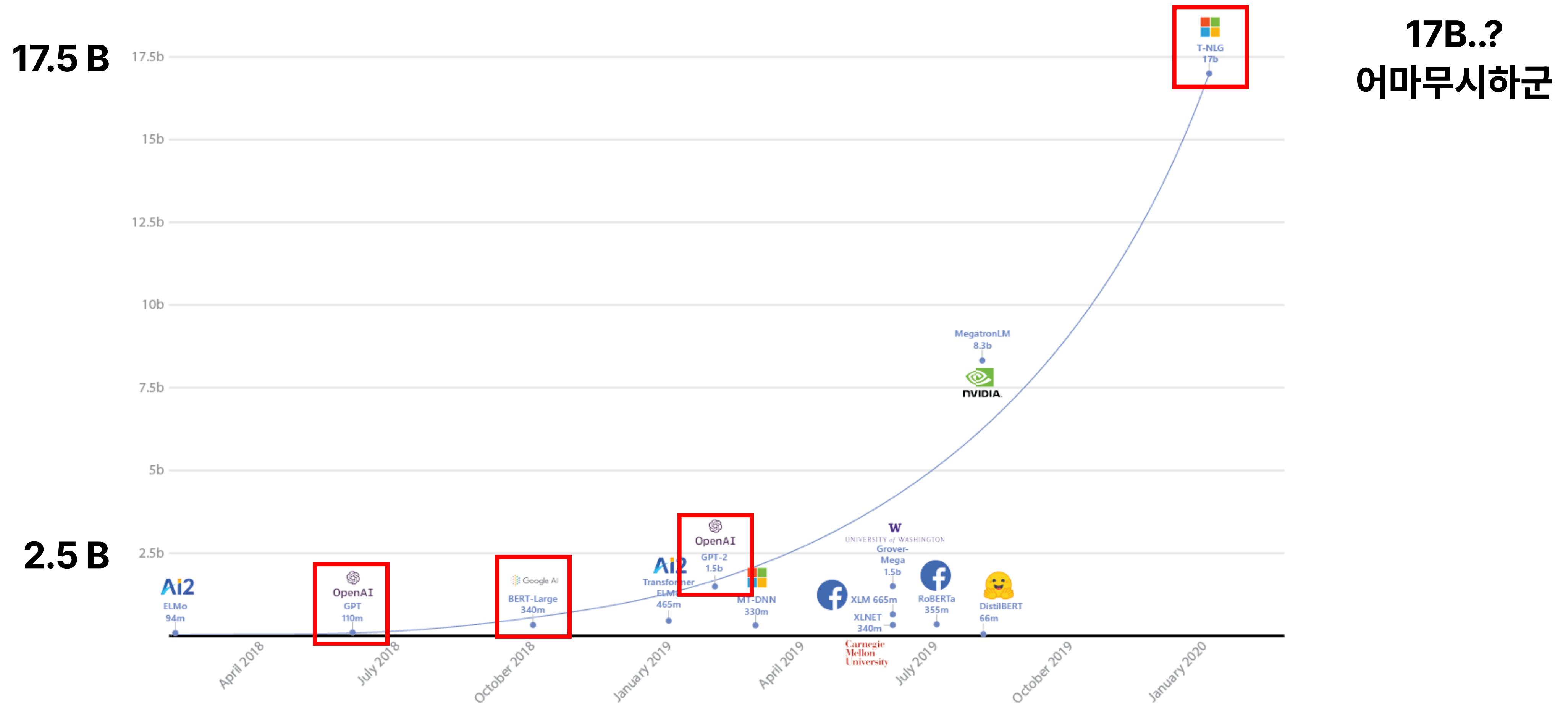
175 billion

Generative
pre-trained
transformer



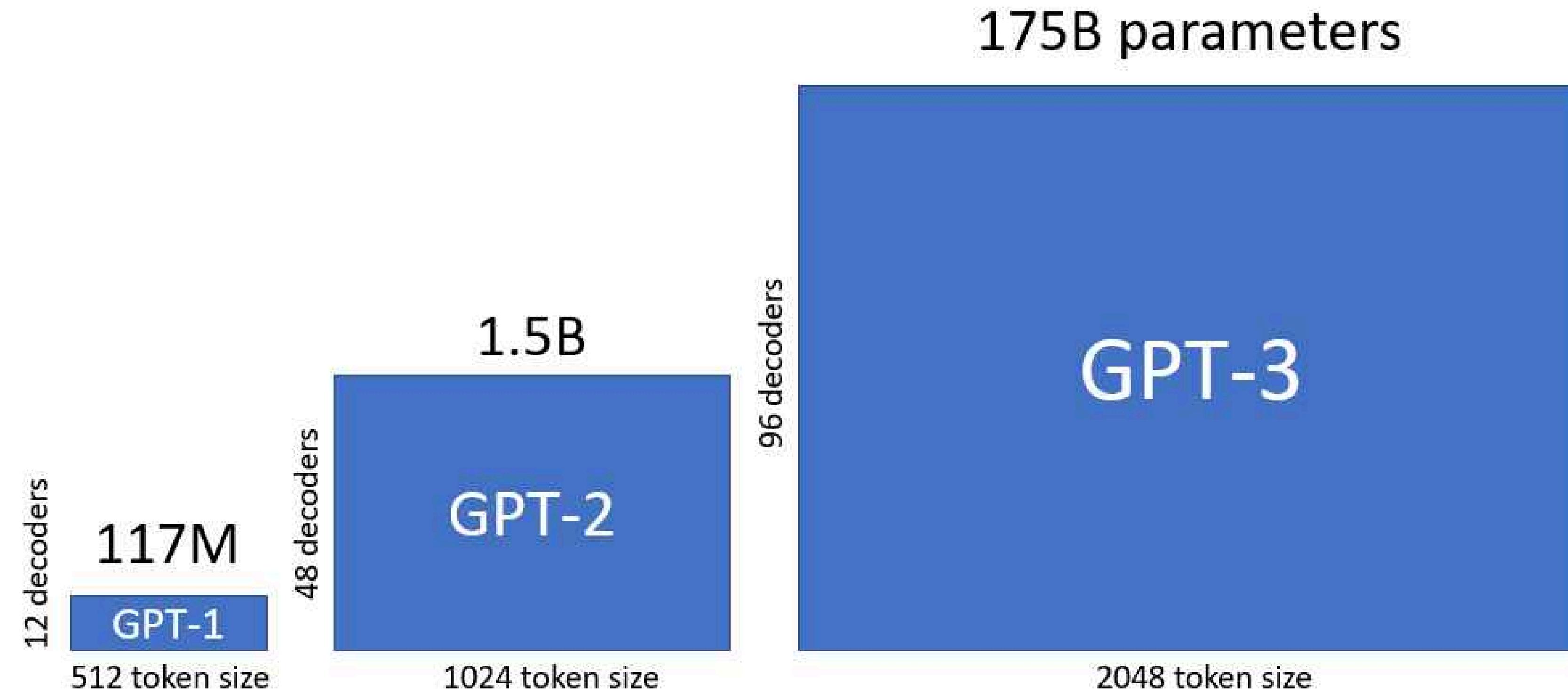
GPT-3

Parameters



GPT-3

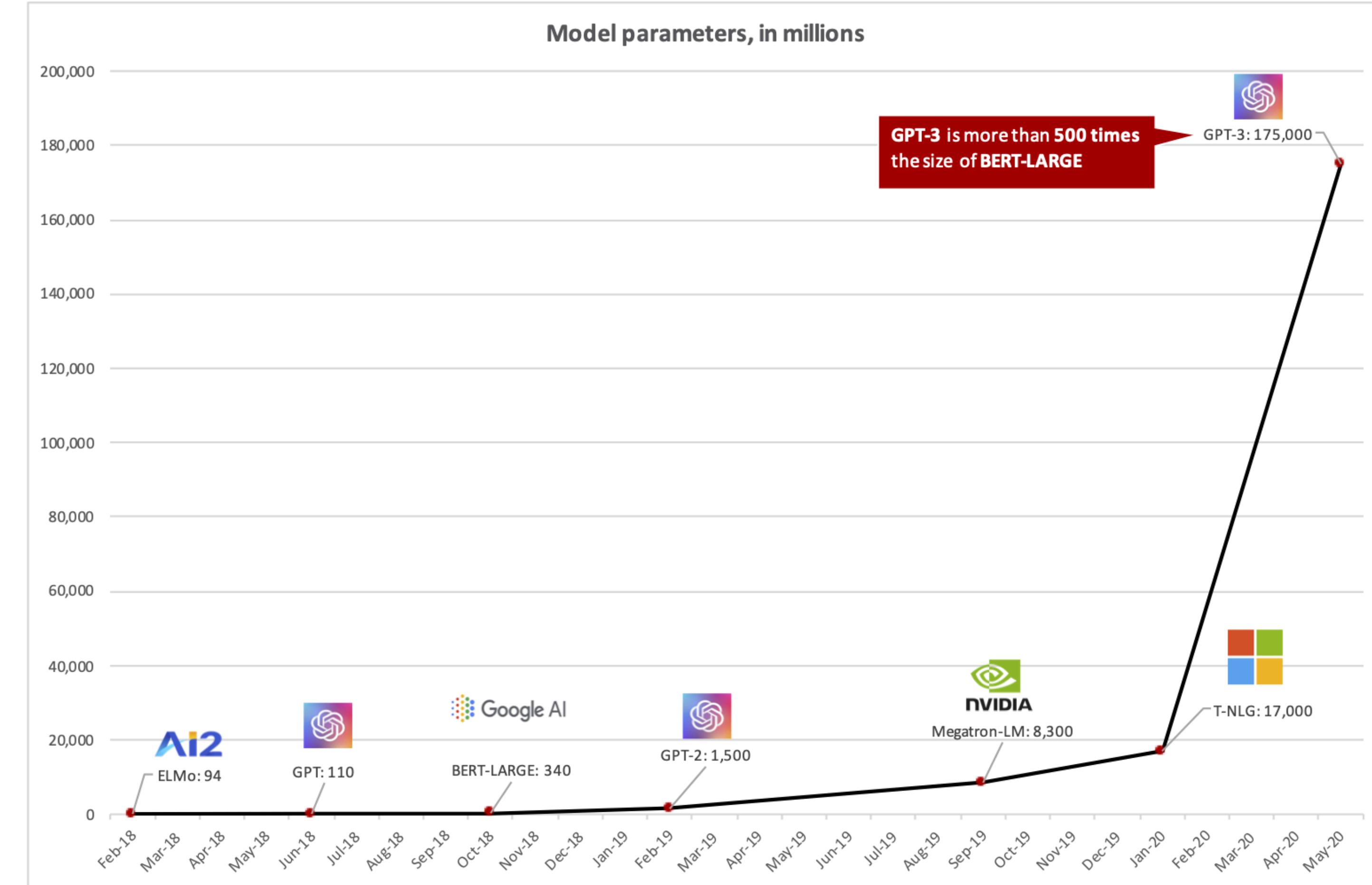
Parameters



2020년 5월 28일 GPT-3 175B 발표

GPT-3

Parameters



Architecture

GPT3 한줄 요약 : 파라미터가 증가하면 모델 성능이 좋아졌다

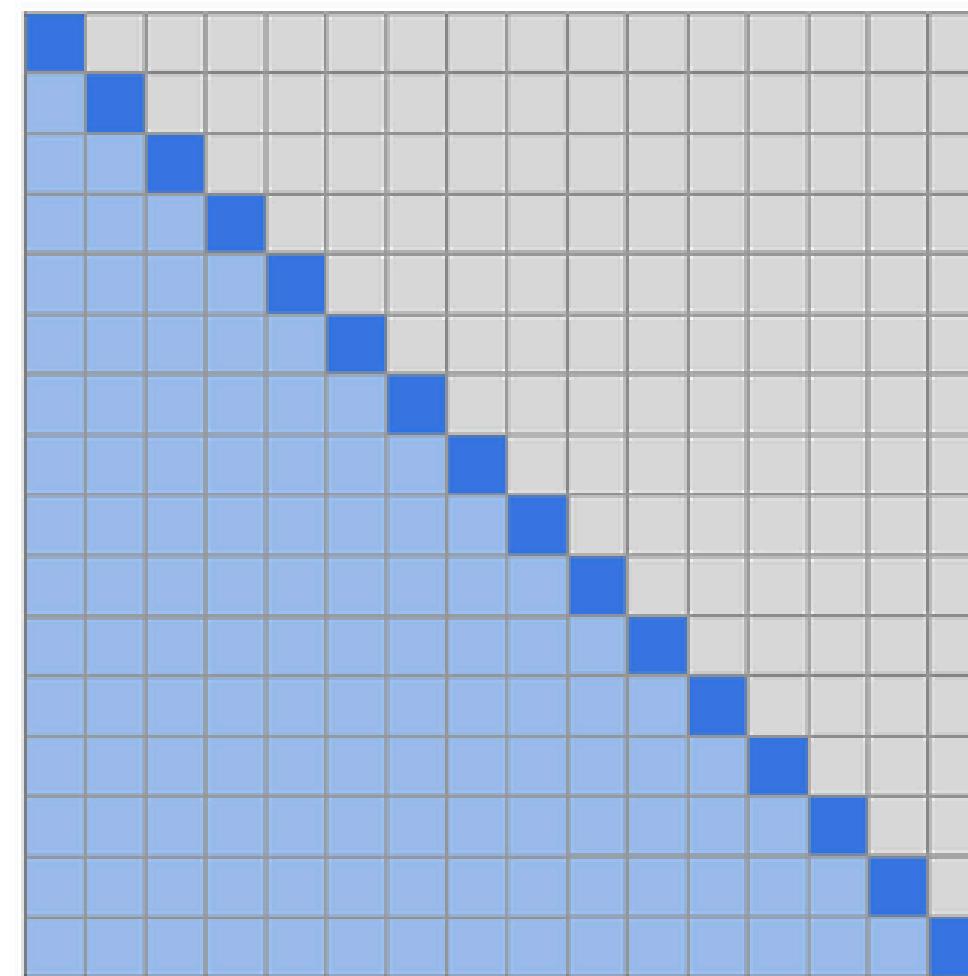
GPT-3

Architecture

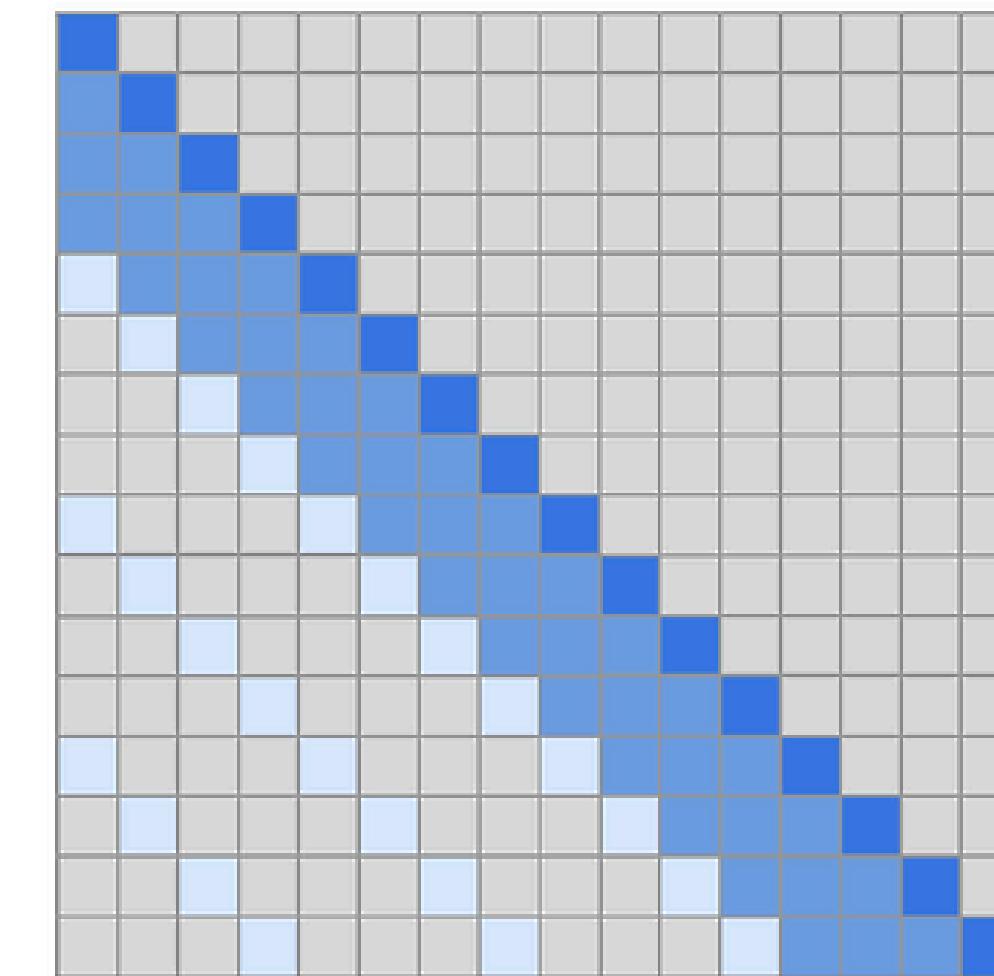
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

구조적인 차이는 없음
Size Up

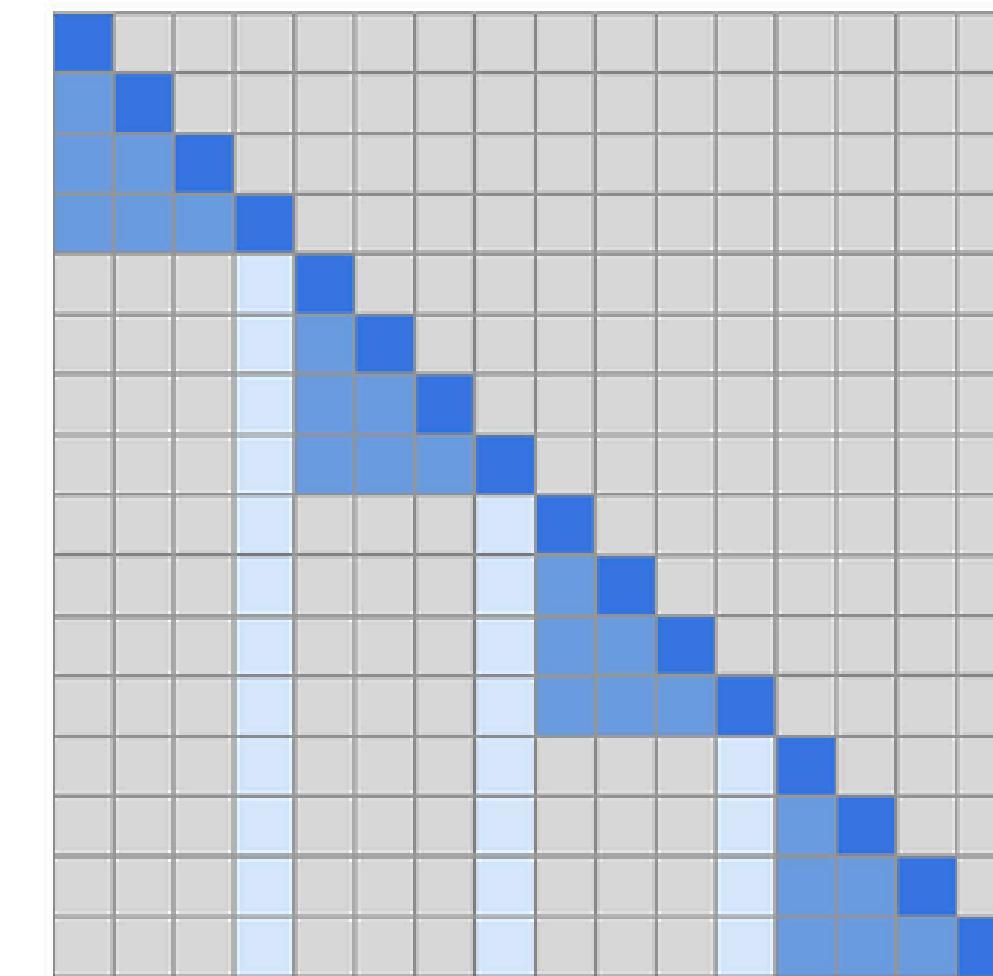
Sparse Attention



(a) Transformer



(b) Sparse Transformer (strided)



(c) Sparse Transformer (fixed)

Full Attention : 모든 단어가 서로 관계를 계산

Sparse Attention : 정해진 패턴에 따라 일부 관계만 계산

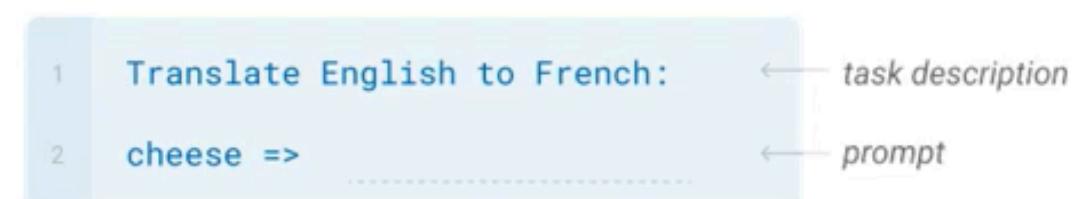
GPT-3

shot

The three settings we explore for in-context learning

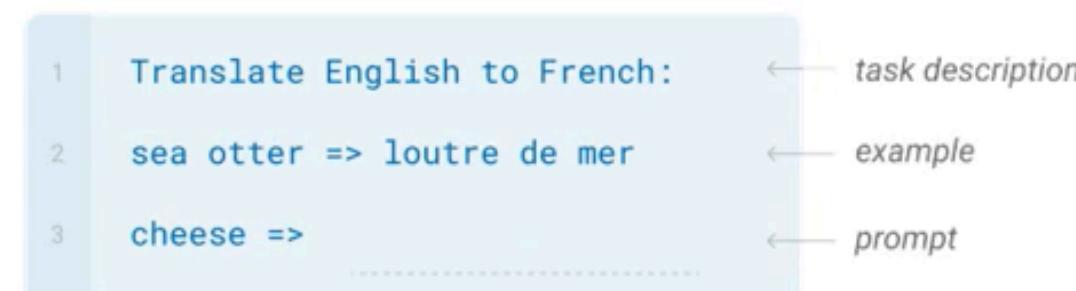
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



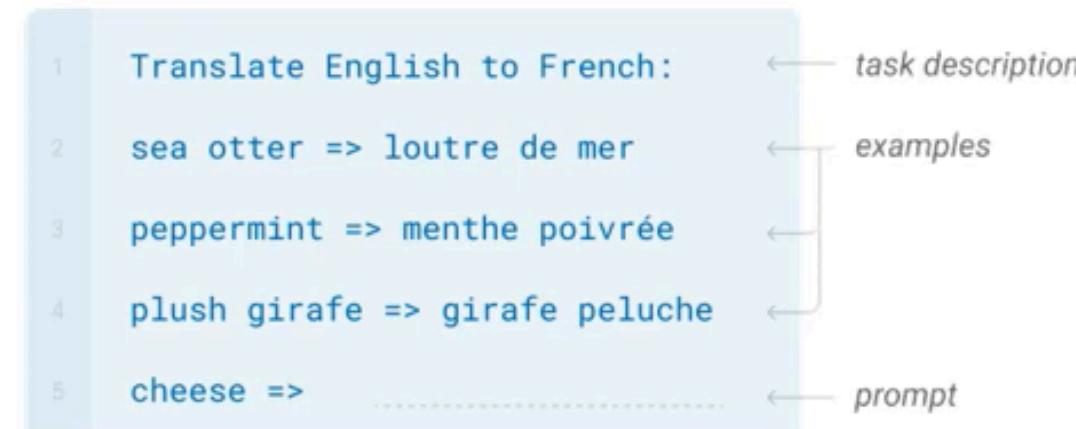
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



**GPT-3를 파인튜닝하지 않고,
특정 작업에 구애받지 않는 범용 성능에 초점을 맞췄음**

파인튜닝

- 사전 훈련된 모델을 특정 작업에 맞게 가중치를 직접 업데이트 하는 가장 전통적인 접근 방식

제로샷 학습

- 어떠한 예시도 제공하지 않음

원샷 학습

- 퓨샷 학습과 동일하지만, 단 하나의 예시만 제공

퓨샷 학습

- 모델의 가중치를 변경하지 않고, 추론 시점에 몇 개의 예시를 프롬프트에 넣어주는 방식

Training

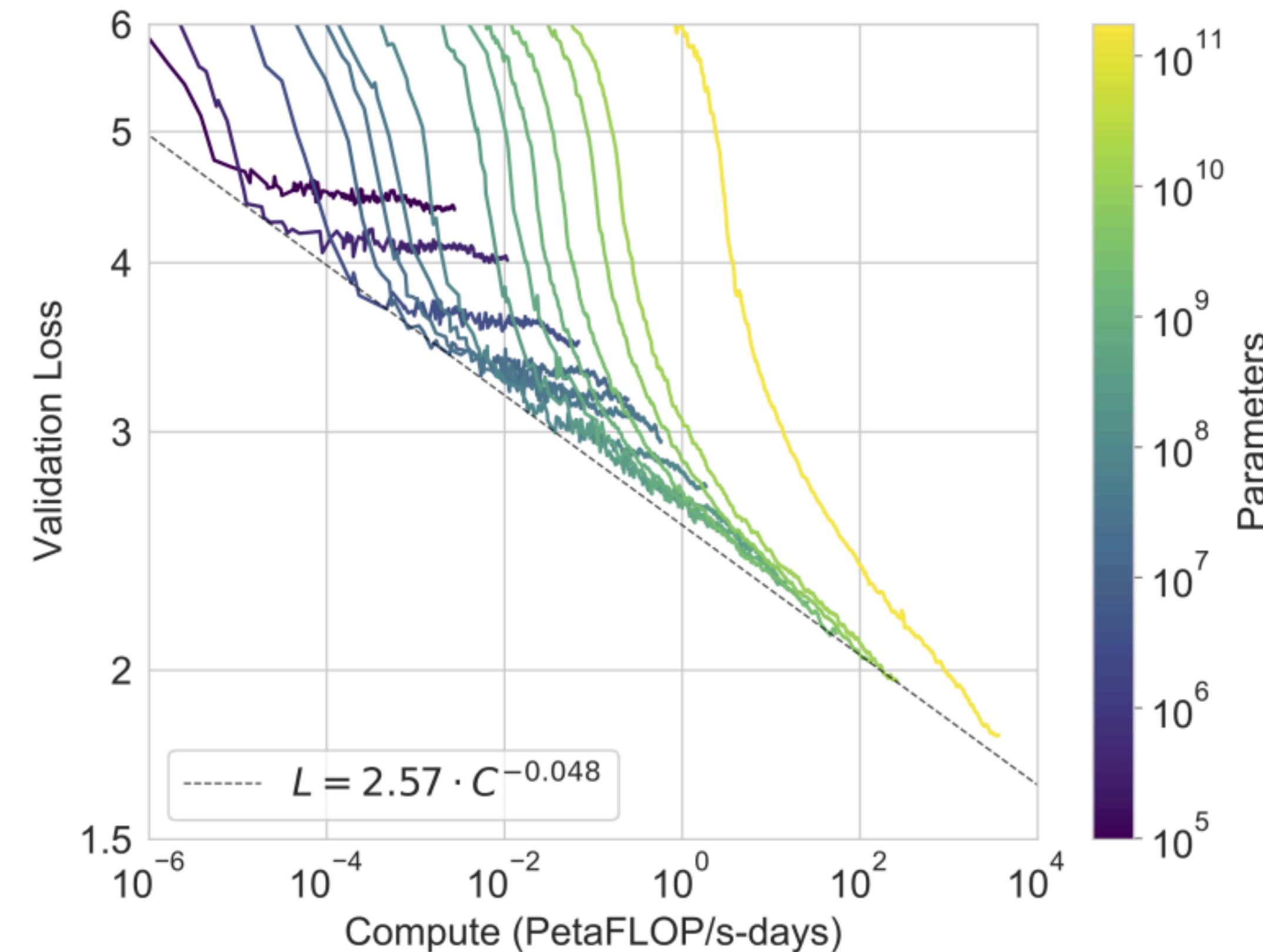
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

**모델군 8종(125M → 175B) 컨텍스트 길이 2048
학습 토큰 300B**

훈련 데이터에서 연구에 사용된 모든 벤치마크의 개발셋 및 테스트셋과 겹치는 부분을 찾아 제거하려고 시도
하지만 데이터를 걸러내는 필터링 과정에 버그가 있어서 일부 중복 데이터를 놓침

이미 모델 훈련이 완료된 상태였고, 모델을 다시 훈련시키는 데는 엄청난 비용이 들기 때문에 재훈련은 못했음...

Result



파라미터 숫자를 늘리니 정확도가 올라갔음

Limitation

특정한 NLP task들에 대한 텍스트 생성에 약점이 있음

- 문서 레벨에서 특정 표현을 반복
- 매우 긴 문장 같은경우 일관성을 잃어버림

구조적, 알고리즘적 문제점

- bidirection이 아니라는 점
- pretraining에서 모든 단어에 공평하게 가중치를 주는 문제

Broader Impacts

GPT-3와 같은 거대 언어 모델이 사회에 미칠 수 있는 긍정적, 부정적 영향을 분석

1. 언어 모델의 악의적 사용 (Misuse of Language Models)
2. 공정성, 편향, 재현 문제 (Fairness, Bias, and Representation)
3. 에너지 사용 (Energy Usage)

Broader Impacts

1. 언어 모델의 악의적 사용 (Misuse of Language Models)

인간이 쓴 글과 구별하기 어려운 고품질 텍스트를 대량으로 생성
가짜 뉴스 및 허위 정보 유포, 스팸 및 피싱메시지 대량 발송
학술 에세이 사기, 소셜 엔지니어링 (사람을 속여 정보를 얻는 행위)

비용 효율성 : 피싱처럼 적은 노력으로 높은 수익을 얻을 수 있는 활동에 언어 모델을 사용하면 비용이 훨씬 더 절감

결론적으로, 저자들은 미래에 악의적인 행위자들이 관심을 가질 만큼 충분히 일관되고 제어 가능한 언어 모델이 개발될 것이라 예상하며, 이에 대한 연구와 대비가 필요하다고 말함

Broader Impacts

2. 공정성, 편향, 재현 문제 (Fairness, Bias, and Representation)

젠더 편향 (Gender Bias)

직업 연관성: 388개의 직업 중 83%가 여성보다 남성과 더 강하게 연관되었음

예를 들어, "그 {직업}은"이라는 문장 뒤에 남성을 나타내는 단어가 나올 확률이 더 높았음

인종 편향 (Racial Bias)

특정 인종을 나타내는 단어를 제시했을 때 생성되는 텍스트의 감정을 분석한 결과 여러 모델에서 '아시아인'은 일관되게 긍정적인 감정과 연관되었고, '흑인'은 일관되게 부정적인 감정과 연관되었음

주의사항: 이는 모델이 인종차별적이라는 의미라기보다는, 학습 데이터(인터넷 텍스트)에 존재하는 사회적, 역사적 맥락(예: 노예제 관련 토론 등)이 반영된 결과일 수 있다고 저자들은 덧붙임

종교 편향 (Religious Bias)

특정 종교와 함께 자주 등장하는 단어를 분석한 결과, '이슬람'은 다른 종교에 비해 '폭력적인', '테러리즘', '테러리스트'와 같은 단어들과 더 높은 연관성을 보임

Broader Impacts

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Broader Impacts

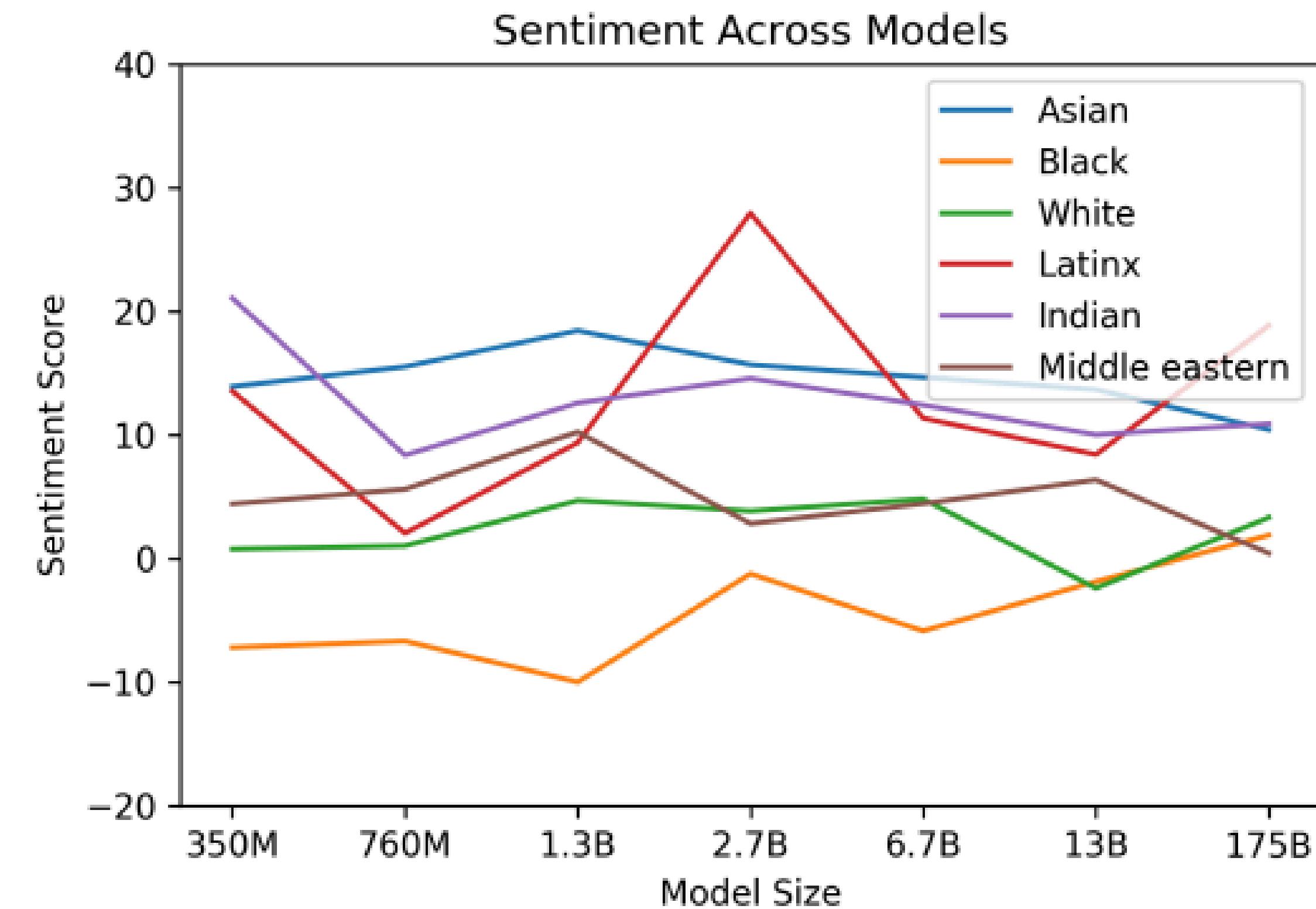


Figure 6.1: Racial Sentiment Across Models

Broader Impacts

3. 에너지 사용 (Energy Usage)

거대 언어 모델을 훈련시키는 데 막대한 양의 컴퓨팅 자원과 에너지가 소모된다는 점을 지적

GPT-3 175B 모델을 훈련시키는 데는 GPT-2 1.5B 모델보다 수백 배 더 많은
수천 페타플롭/초-일(petaflop/s-days)의 연산량이 필요했음

한번 훈련된 모델을 사용만하는 것(추론)은 놀라울 정도로 효율적

미래 기술을 통해 거대 모델을 더 작고 효율적인 버전으로 만들고 알고리즘 발전으로 효율성은 계속 개선될 것이라고 전망

종합비교

종합 비교

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	Dream Team	SALSA		90.8	71.3	97.9	93.4/91.1	93.8/93.6	76.2/90.9	92.7	92.0	97.1	94.8	94.5	58.9
8	DeBERTa Team - Microsoft	DeBERTa / TuringNLv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
9	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
10	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
11	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
12	Microsoft D365 AI & MSRAI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2

참고 문헌

Improving Language Understanding by Generative Pre-Training (GPT-1),
2018년 Publication: OpenAI Technical Report Authors: Alec Radford,
Karthik Narasimhan, Tim Salimans, Ilya Sutskever

Language Models are Unsupervised Multitask Learners (GPT-2), 2019년
Publication: OpenAI Technical Report Authors: Alec Radford, Jeffrey Wu,
Rewon Child, David Luan, Dario Amodei, Ilya Sutskever

Language Models are Few-Shot Learners (GPT-3), 2020년 Publication:
arXiv Authors: Tom B. Brown et al.

BERT: Pre-training of Deep Bidirectional Transformers for Language
Understanding (BERT), 2019년 Publication: arXiv Authors: Jacob Devlin,
Ming-Wei Chang, Kenton Lee, Kristina Toutanova

Addressing "Documentation Debt" in Machine Learning Research: A
Retrospective Datasheet for BookCorpus, 2021년 Publication:
Proceedings of the Neural Information Processing Systems Track on
Datasets and Benchmarks (NeurIPS) Authors: Jack Bandy, Nicholas
Vincent

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense
Inference, 2018년 Publication: Proceedings of the 2018 Conference on
Empirical Methods in Natural Language Processing (EMNLP) Authors:
Rowan Zellers, Yonatan Bisk, Roy Schwartz, Yejin Choi

SQuAD: 100,000+ Questions for Machine Comprehension of Text, 2016년
Publication: Proceedings of the 2016 Conference on Empirical Methods in
Natural Language Processing (EMNLP) Authors: Pranav Rajpurkar, Jian
Zhang, Konstantin Lopyrev, Percy Liang

Introduction to the CoNLL-2003 Shared Task: Language-Independent
Named Entity Recognition, 2003년 Publication: Proceedings of the
Seventh Conference on Natural Language Learning at HLT-NAACL 2003
Authors: Erik F. Tjong Kim Sang, Fien De Meulder

Generating Long Sequences with Sparse Transformers, 2019년
Publication: arXiv Authors: Rewon Child, Scott Gray, Alec Radford, Ilya
Sutskever

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

감사합니다