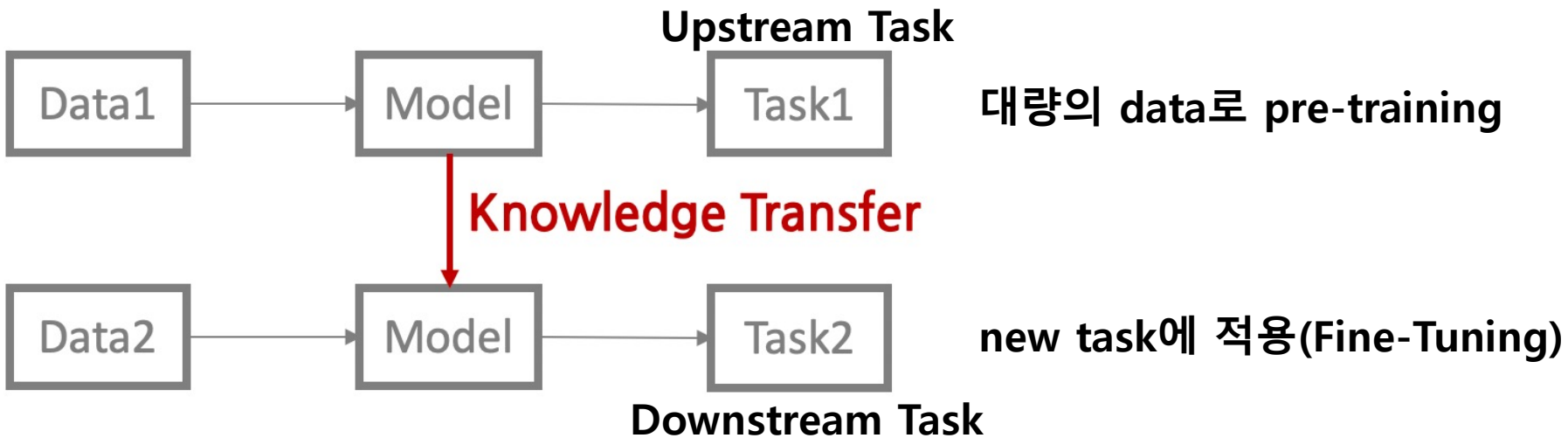
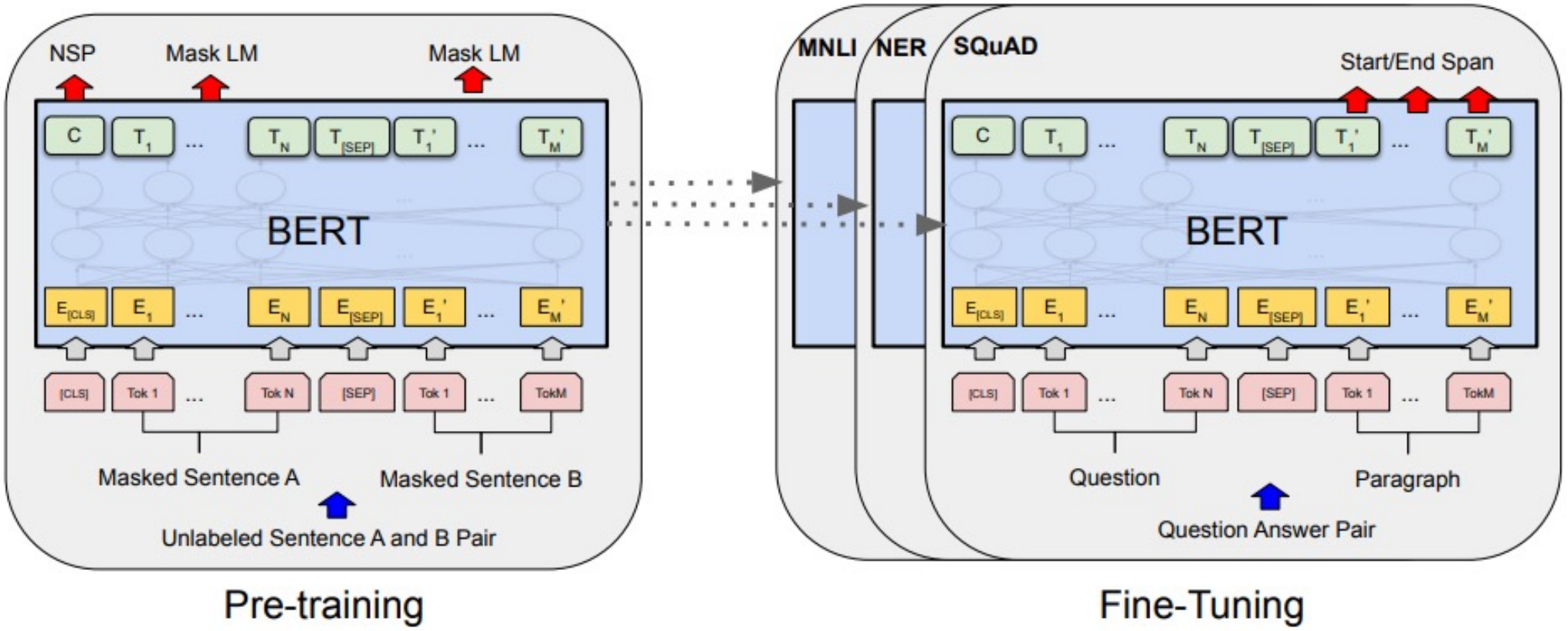
T5 : Text-To-Text Transfer Transformer

Transfer Learning



GPT, ULMFit, BERT, XLNet, RoBERTa, ALBERT, Reformer, MT-DNN...

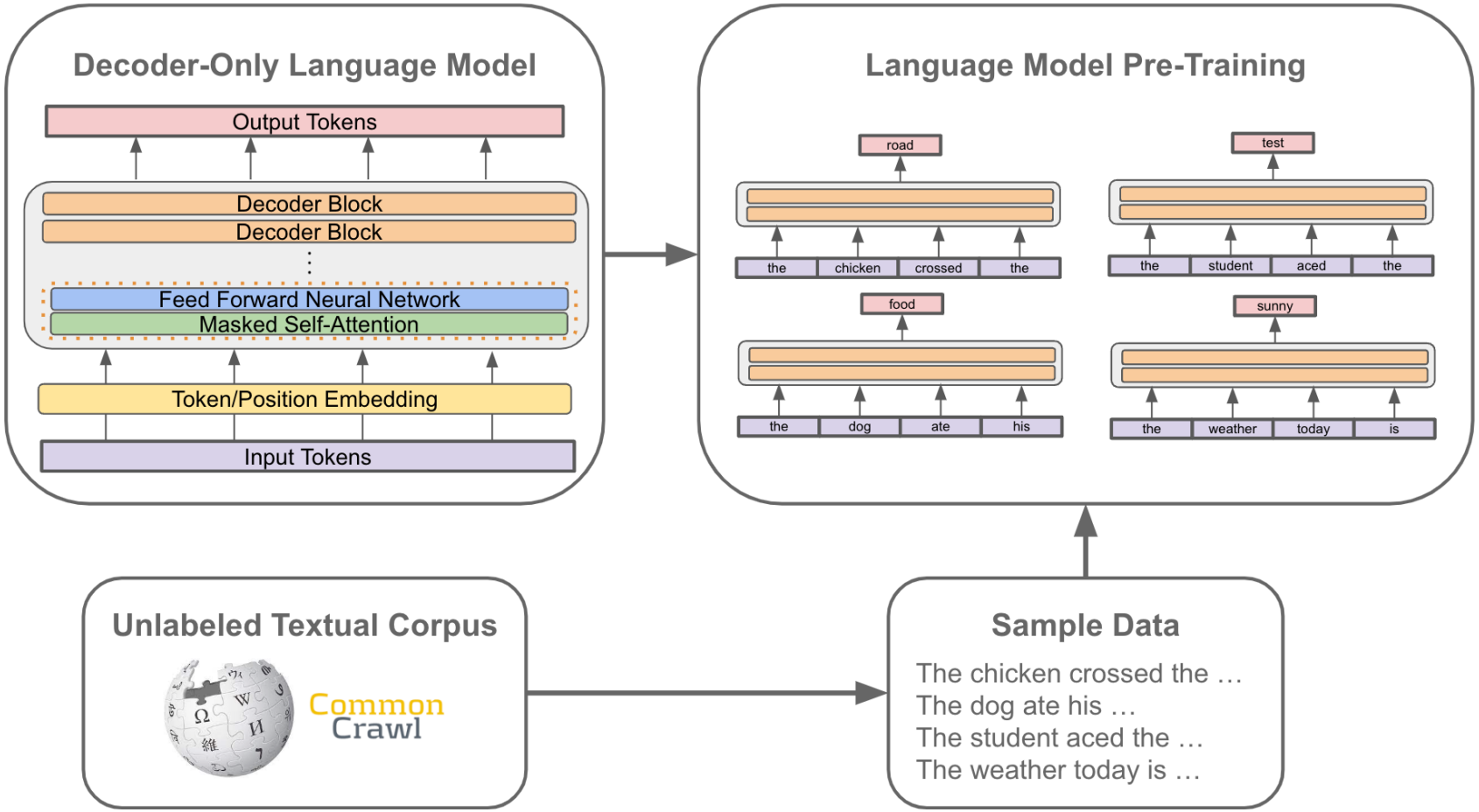
BERT



Pre-training : 마스킹 단어 예측(MLM), 문장 순서 파악(NSP)

New task : 텍스트 분류(MNLI), 개체명 인식(NER), 질문 응답(SQuAD)

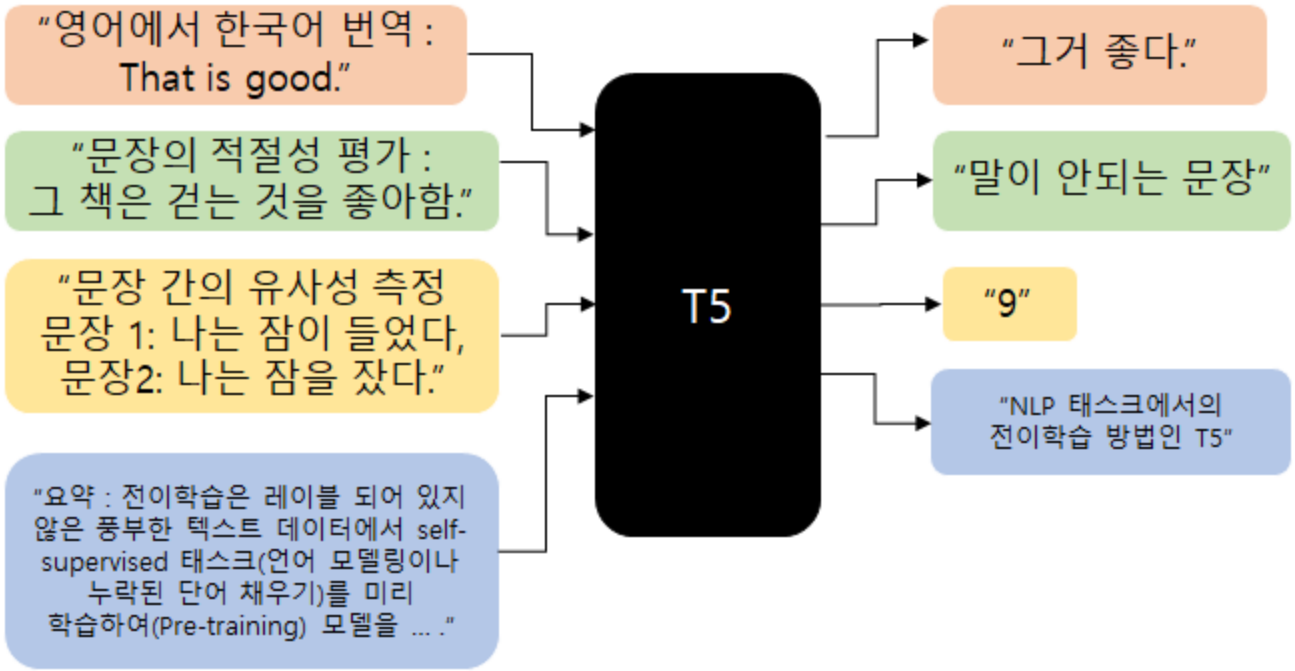
GPT



Pre-training : 주어진 텍스트의 다음 단어 예측(Language Modeling)

New task : 텍스트 생성, 스토리텔링, 번역

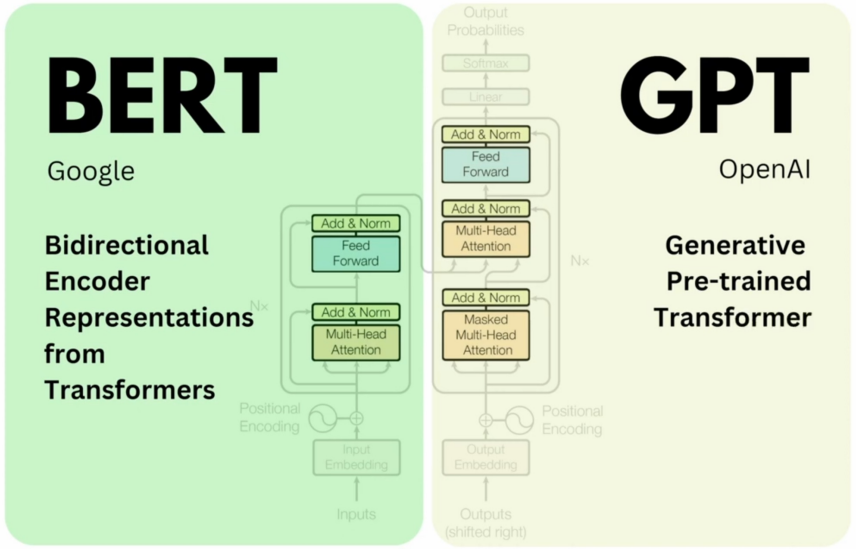
T5


"영어에서 한국어 번역 : That is good."
"문장의 적절성 평가 : 그 책은 걷는 것을 좋아함."
"문장 간의 유사성 측정 문장 1: 나는 잠이 들었다, 문장2: 나는 잠을 잤다."
"요약 : 전이학습은 레이블 되어 있지 않은 풍부한 텍스트 데이터에서 self-supervised 태스크(언어 모델링이나 누락된 단어 채우기)를 미리 학습하여(Pre-training) 모델을 ... ."
T5
"그거 좋다."
"말이 안되는 문장"
"9"
"NLP 태스크에서의 전이학습 방법인 T5"

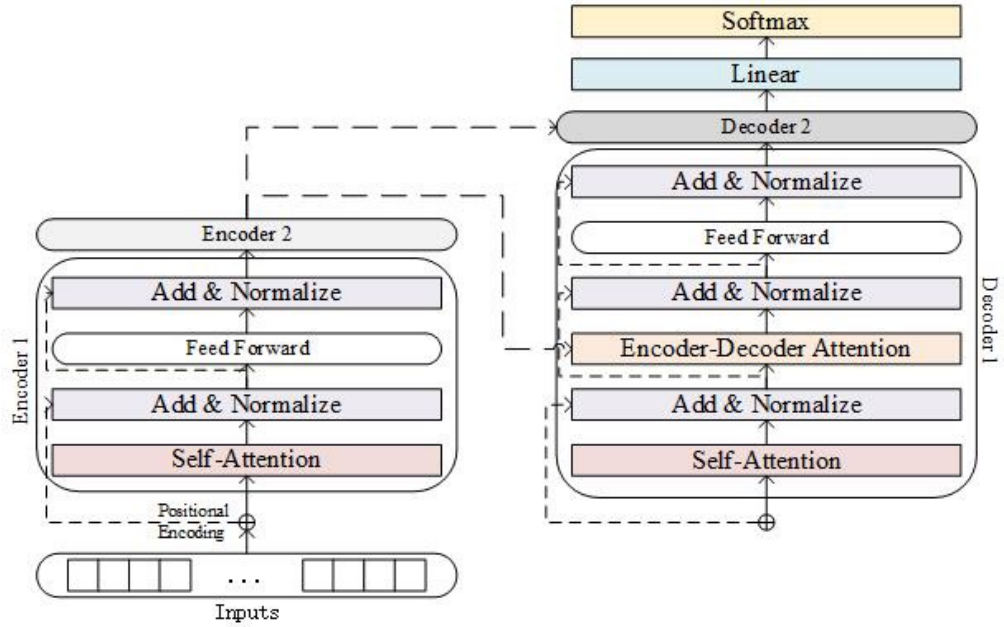
T5

1. Encoder-Decoder
2. Relative Positional Embedding
3. Input & Output Format
4. The Colossal Clean Crawled Corpus (C4)

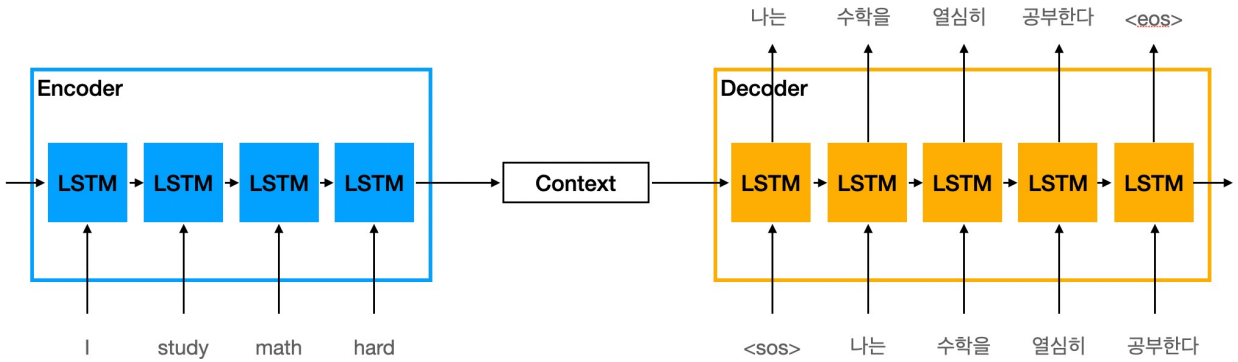
1. Encoder-Decoder



Vs



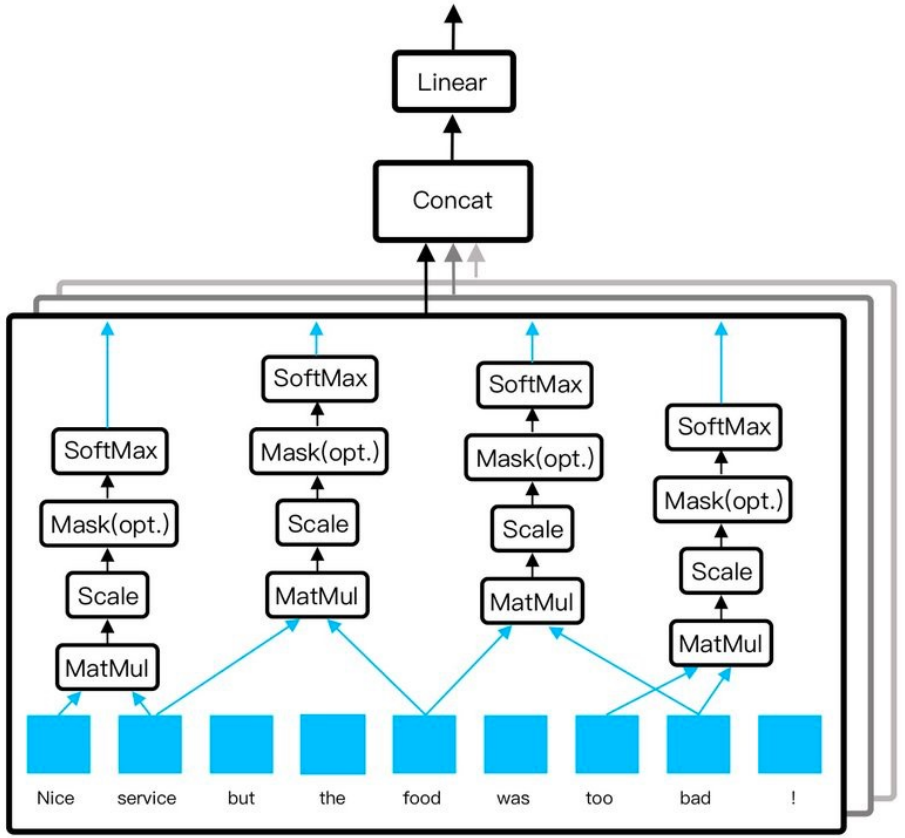
2. Relative Positional Embedding



서비스는 좋았지만 음식이 별로였다!

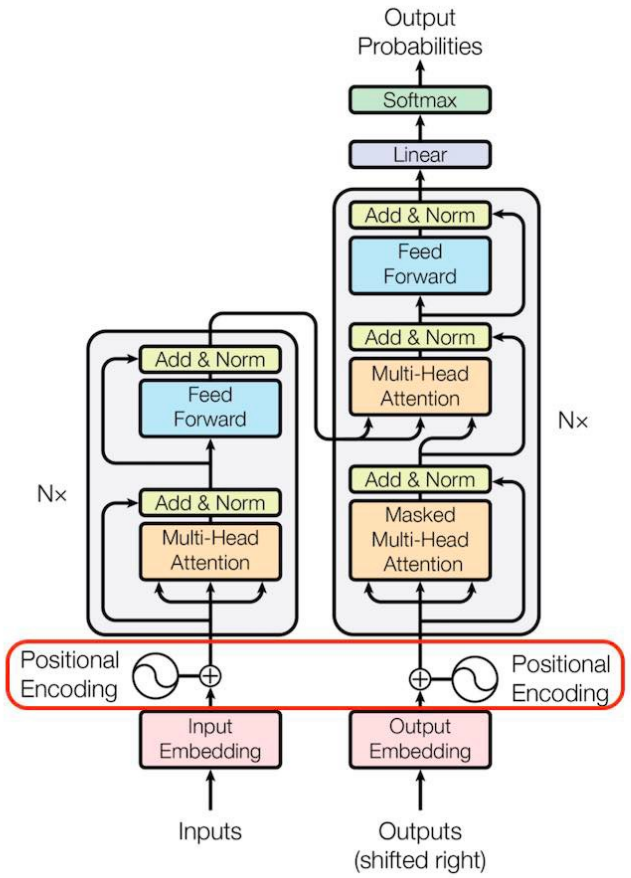
음식은 좋았지만 서비스가 별로였다?

서비스는 음식이었지만 별로는 좋았다?

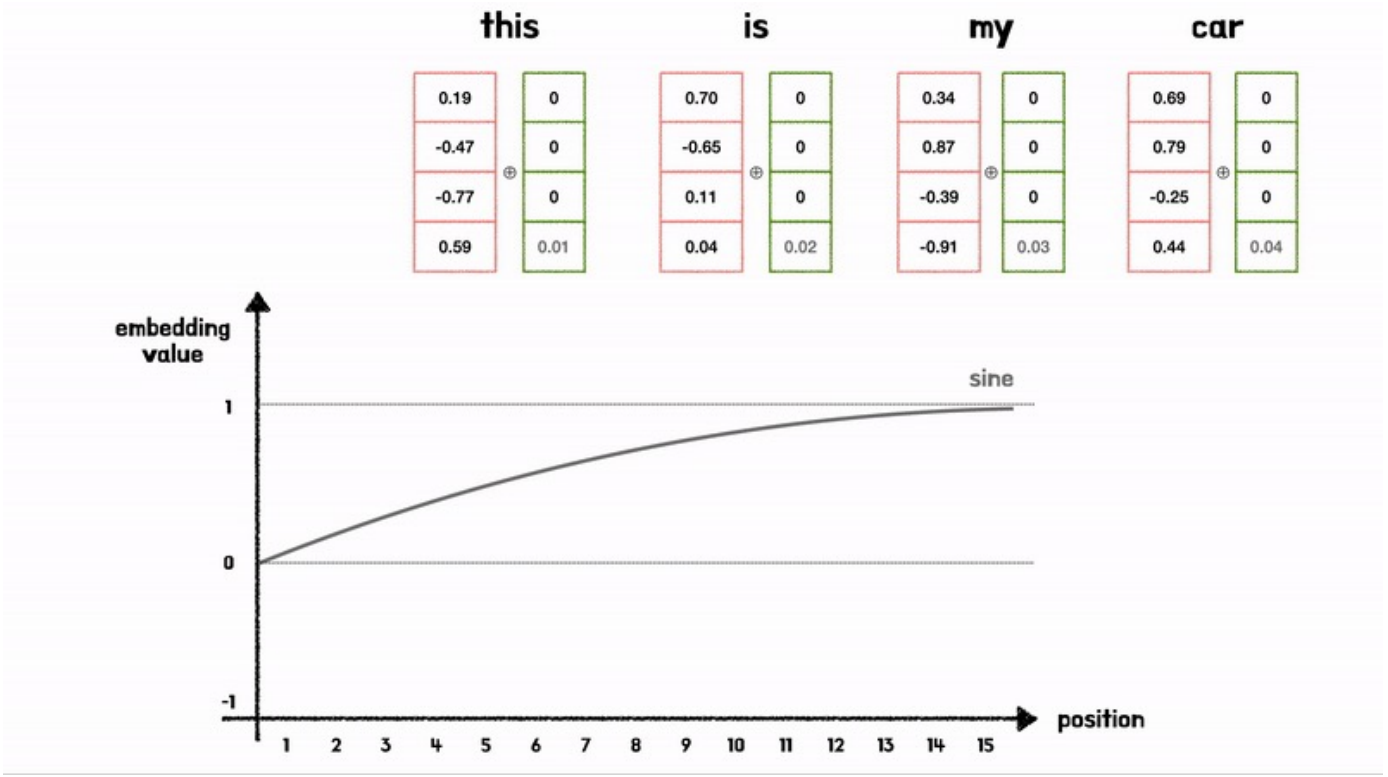


Transformer : Self-Attention

2. Relative Positional Embedding

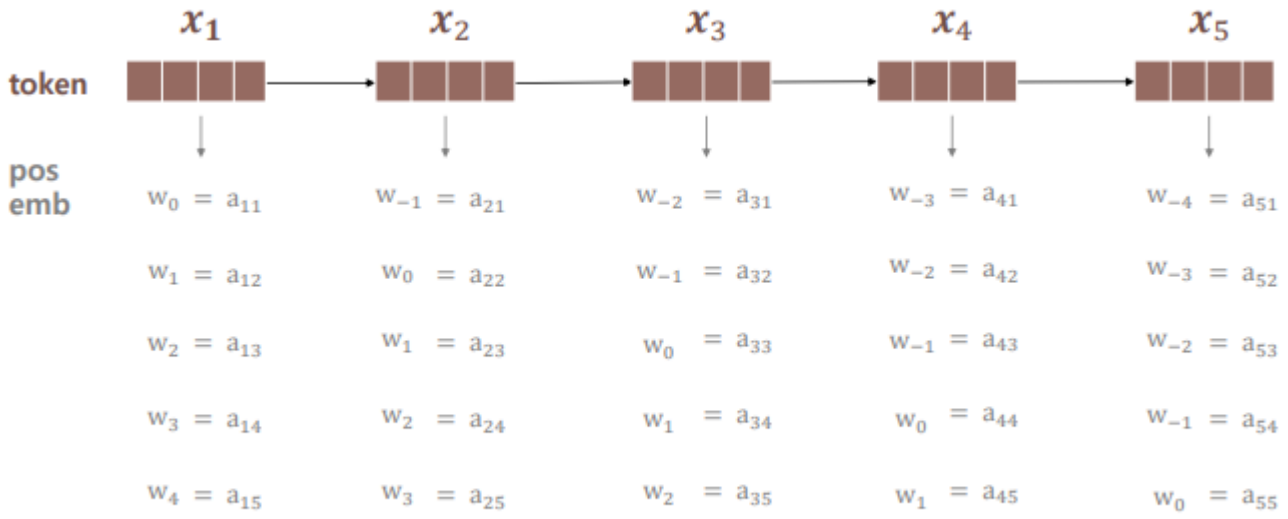


Architecture of Transformer



- 위치 정보를 표현할 n차원 Embedding 벡터를 생성
- 벡터의 짝수 index dimension에 sin 함수를 대입
- 벡터의 홀수 index dimension에 cos 함수를 대입
- Frequency를 다르게 함으로써 아주 큰 값도 작은 수로 표현 가능

2. Relative Positional Embedding



- 문장을 구성하는 단어 수 만큼의 차원을 갖는 Embedding 벡터를 생성
- 단어와, 그 단어의 주변 단어들 간의 위치 정보를 계산하여 해당하는 자리에 대입

- <u>문장 길이가 길어져도 단어들 간 위치 파악이 쉬움</u>

3. Input & Output Format

기존 Format

Format : Input-Output의 형식

Input	target
그는 한번도 여기에 왔다	0 (integer)
한 남자가 음식을 먹고 있다 한 남자가 뭔가를 먹고 있다	4.2 (integer)
What's your name?	Wie heißen sie?

Text-to-Text Format

Format : Input-Output의 형식

문장의 적절성 평가: 그는 한번도 여기에 왔다

→ 적합하지 않음

문장 간 유사도 측정 문장 1: 한 남자가 음식을 먹고 있다

문장 2: 한 남자가 뭔가를 먹고 있다

→ "4.2" (string)

translate English to German: What's your name?

→ Wie heißen sie?

3. Input & Output Format

sentence1	sentence2	score
A plane is taking off.	An air plane is taking off.	5.000
A man is playing a large flute.	A man is playing a flute.	3.800
A man is spreading shreded cheese on a pizza.	A man is spreading shredded cheese on an uncooked pizza.	3.800
Three men are playing chess.	Two men are playing chess.	2.600
A man is playing the cello.	A man seated is playing the cello.	4.250
Some men are fighting.	Two men are fighting.	4.250
A man is smoking.	A man is skating.	0.500
The man is playing the piano.	The man is playing the guitar.	1.600
A man is playing on a guitar and singing.	A woman is playing an acoustic guitar and singing.	2.200
A person is throwing a cat on to the ceiling.	A person throws a cat on the ceiling.	5.000
The man hit the other man with a stick.	The man spanked the other man with a stick.	4.200
A woman picks up and holds a baby kangaroo.	A woman picks up and holds a baby kangaroo in her arms.	4.600
A man is playing a flute.	A man is playing a bamboo flute.	3.867
A person is folding a piece of paper.	Someone is folding a piece of paper.	4.667
A man is running on the road.	A panda dog is running on the road.	1.667

두 문장의 의미가 같은지 다른지 유사도를 판단하는 “STS-B” task

3. Input & Output Format

4. The Colossal Clean Crawled Corpus (C4)

- Common Crawl dataset으로부터 수집된 대량의 text data를 전처리
- .,?,! 등과 같이 종결형 문장 부호로 끝나는 문장들만 남긴 후 나머지 문장 삭제
- 5문장 미만으로 구성된 page는 폐기하고, 최소 3단어 이상으로 구성된 문장만 남김
- "List of Dirty, Naughty, Obscene or Otherwise Bad Words"에 포함된 비속어 및 은어와 같은 단어들 삭제
- Javascript 단어가 있는 문장 삭제
- "lorem ipsum" 구문이 보이는 페이지 삭제
- {} 중괄호를 포함하고 있는 페이지 삭제
- 3문장 이상으로 구성된 span이 2번 이상 반복되면 해당 페이지 삭제

Experiments

- **Expt.1 : Model Structures**
- **Expt.2 : Unsupervised Objectives**
- **Expt.3 : Pre-training Dataset**
- **Expt.4 : Fine-tuning method**
- **Expt.5 : Scaling**

Experiments – Benchmark

- GLUE & SuperGLUE(SGLUE)
- CNN/Daily Mail(CNNDM) : 텍스트 요약
- SQuAD : 질의응답(독해력 체크)
- EnDe : English to Deutsch
- EnFr : English to French
- EnRO : English to Russian

Experiments – Benchmark

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical? 문법적 오류	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral? 감정 분석	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A? 문장 유사도	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	문장 유사도 How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	문장 유사도 Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B? 자연어 추론	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	자연어 추론 Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	자연어 추론 Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun? 공지시 해결	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

GLUE Dataset

Expt.1 Model Structures

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39
Encoder-decoder	LM	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	LM	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	LM	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	LM	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

Table 2: Performance of the different architectural variants described in Section 3.2.2. We use P to refer to the number of parameters in a 12-layer base Transformer layer stack and M to refer to the FLOPs required to process a sequence using the encoder-decoder model. We evaluate each architectural variant using a denoising objective (described in Section 3.1.4) and an autoregressive objective (as is commonly used to train language models).

Expt.2 : Unsupervised Objectives

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

Table 3: Examples of inputs and targets produced by some of the unsupervised objectives we consider applied to the input text "Thank you for inviting me to your party last week ." Note that all of our objectives process *tokenized* text. For this particular sentence, all words were mapped to a single token by our vocabulary. We write *(original text)* as a target to denote that the model is tasked with reconstructing the entire input text. <M> denotes a shared mask token and <X>, <Y>, and <Z> denote sentinel tokens that are assigned unique token IDs. The BERT-style objective (second row) includes a corruption where some tokens are replaced by a random token ID; we show this via the greyed-out word apple.

Expt.2 : Unsupervised Objectives

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Prefix language modeling	80.69	18.94	77.99	65.27	**26.86**	39.73	**27.49**
BERT-style (Devlin et al., 2018)	**82.96**	**19.17**	**80.65**	**69.85**	**26.78**	**40.03**	**27.41**
Deshuffling	73.17	18.59	67.61	58.47	26.11	39.30	25.62

Table 4: Performance of the three disparate pre-training objectives described in Section 3.3.1.

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
BERT-style (Devlin et al., 2018)	82.96	19.17	**80.65**	69.85	26.78	**40.03**	27.41
MASS-style (Song et al., 2019)	82.32	19.16	80.10	69.28	26.79	**39.89**	27.55
★ Replace corrupted spans	83.28	**19.24**	**80.88**	**71.36**	**26.98**	39.82	**27.65**
Drop corrupted tokens	**84.44**	**19.31**	**80.52**	68.67	**27.07**	39.76	**27.82**

Table 5: Comparison of variants of the BERT-style pre-training objective. In the first two variants, the model is trained to reconstruct the original uncorrupted text segment. In the latter two, the model only predicts the sequence of corrupted tokens.

Expt.2 : Unsupervised Objectives

Span length	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (i.i.d.)	**83.28**	19.24	80.88	71.36	**26.98**	**39.82**	**27.65**
2	**83.54**	19.39	**82.09**	**72.20**	**26.76**	**39.99**	**27.63**
3	**83.49**	**19.62**	**81.84**	**72.53**	**26.86**	39.65	**27.62**
5	**83.40**	19.24	**82.05**	**72.23**	**26.88**	39.40	**27.53**
10	82.85	19.33	**81.84**	70.44	**26.79**	39.49	**27.69**

Table 7: Performance of the span-corruption objective (inspired by Joshi et al. (2019)) for different average span lengths. In all cases, we corrupt 15% of the original text sequence.

Expt.3 : Pre-training Dataset

Data set	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Table 8: Performance resulting from pre-training on different data sets. The first four variants are based on our new C4 data set.

Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full data set	0	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**80.97**	**72.03**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	79.78	69.97	**27.02**	**39.71**	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

Table 9: Measuring the effect of repeating data during pre-training. In these experiments, we only use the first N tokens from C4 (with varying values of N shown in the first column) but still pre-train over 2^{35} tokens. This results in the data set being repeated over the course of pre-training (with the number of repeats for each experiment shown in the second column), which may result in memorization (see Figure 6).

Expt.4 : Fine-tuning method

Fine-tuning method	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ All parameters	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Adapter layers, $d = 32$	80.52	15.08	79.32	60.40	13.84	17.88	15.54
Adapter layers, $d = 128$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Adapter layers, $d = 512$	81.54	17.78	79.18	64.30	23.45	33.98	25.81
Adapter layers, $d = 2048$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Gradual unfreezing	82.50	18.95	79.17	**70.79**	26.71	39.02	26.93

Table 10: Comparison of different alternative fine-tuning methods that only update a subset of the model's parameters. For adapter layers, d refers to the inner dimensionality of the adapters.

Expt.5：Scaling

Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	**86.18**	19.66	**84.18**	77.18	27.52	**41.03**	28.19
4× size, 1× training steps	**85.91**	19.73	**83.86**	**78.04**	27.47	40.71	28.10
4× ensembled	84.77	**20.10**	83.09	71.74	**28.05**	40.53	**28.57**
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09

Table 13: Comparison of different methods of scaling up our baseline model. All methods except ensembling fine-tuned models use 4× the computation as the baseline. "Size" refers to the number of parameters in the model and "training time" refers to the number of steps used for both pre-training and fine-tuning.