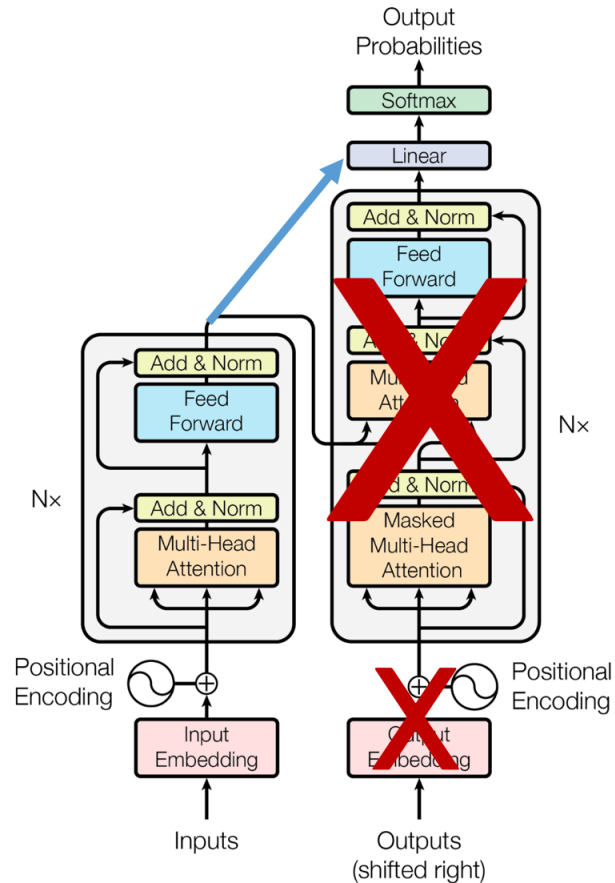
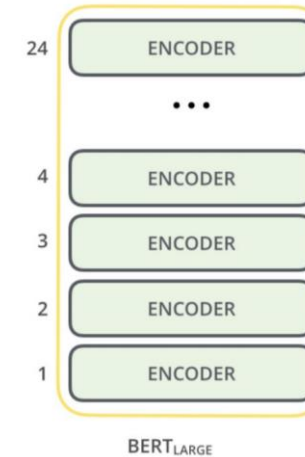
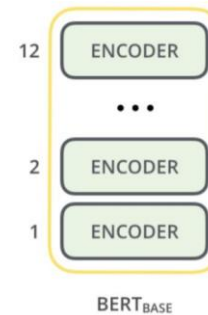


BERT 모델 개요



Bidirectional Encoder Representations from Transformers

Transformer 모델의 인코더 부분을 쌓아 만든 모델



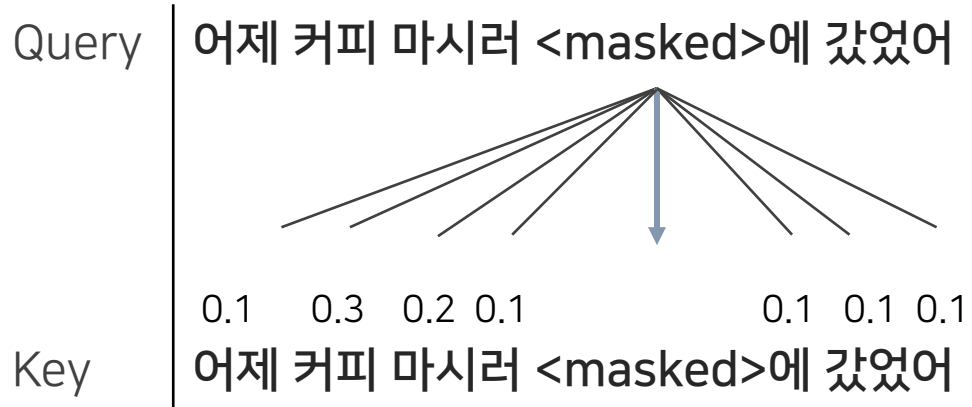
https://ratsgo.github.io/nlpbook/docs/language_model/bert_gpt/

BERT is basically a trained Transformer Encoder stack. (Alammar, 2021)

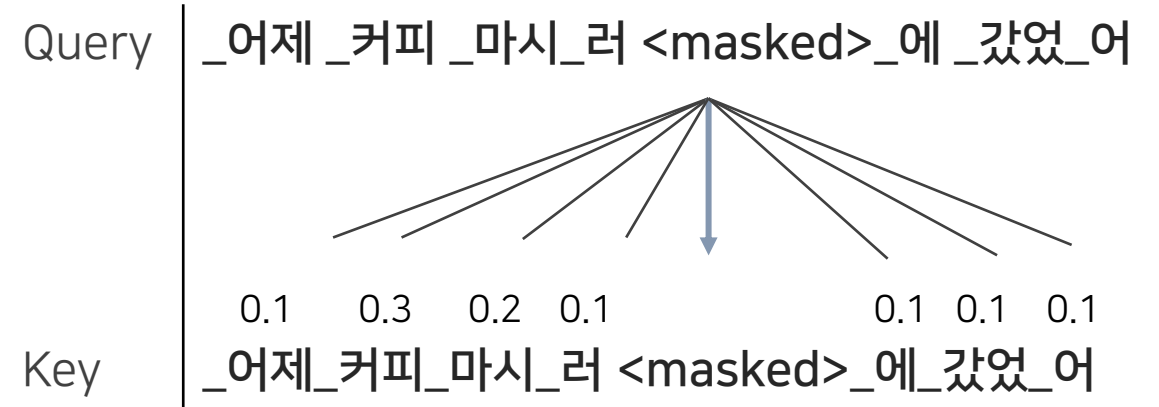
BERT 모델 개요

Self-Attention

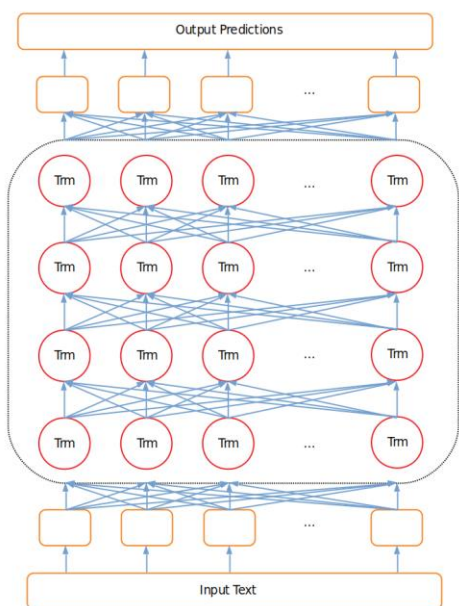
Input text



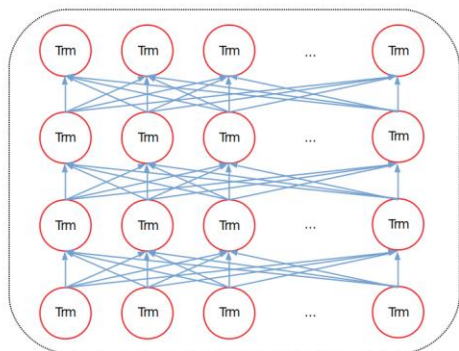
WordPiece embedding



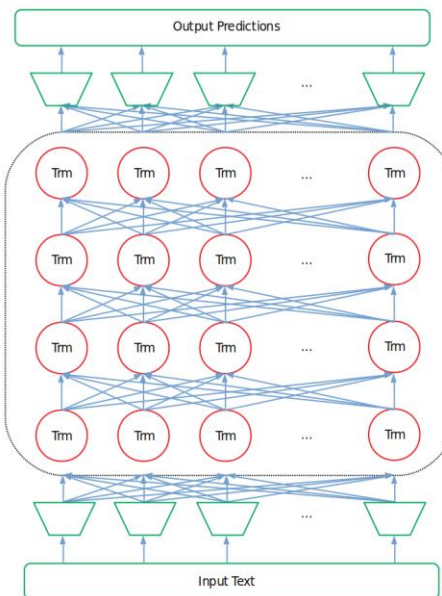
Masked word 를 기준으로 앞뒤 문맥의 단어를 활용해 모델 학습 -> 각 임베딩 토큰의 의미를 잘 담을 수 있음



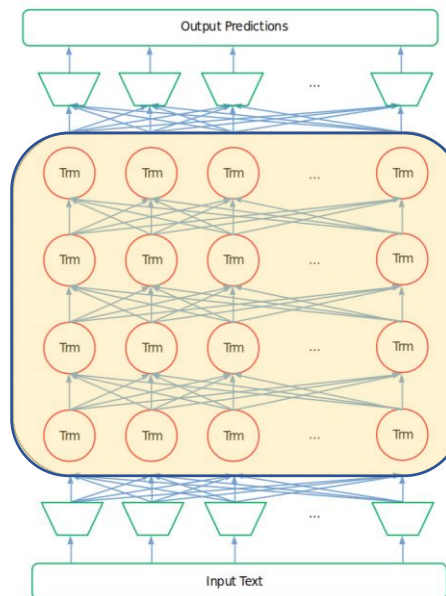
사전 학습된 BERT 모델



입력층 및 출력층 제거

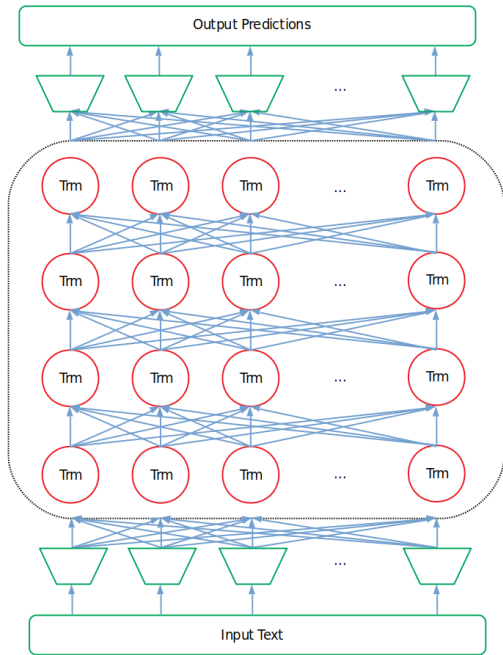


목적에 따라 입력층과 출력층이 수정된 BERT 모델



수정된 입력층과 출력층을 재학습시켜 Fine-Tuning된 BERT 모델

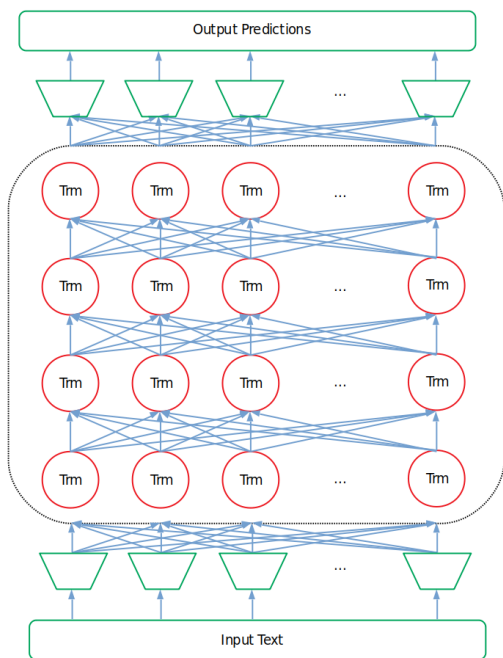
BERT 모델 개요



Output Predictions : NLP Task의 Output Prediction

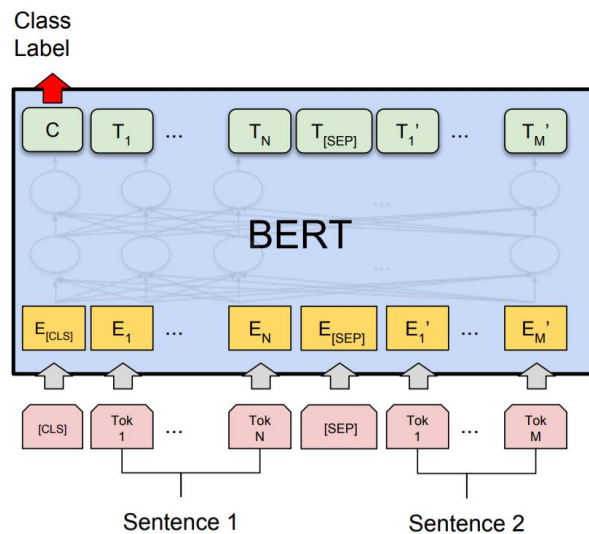
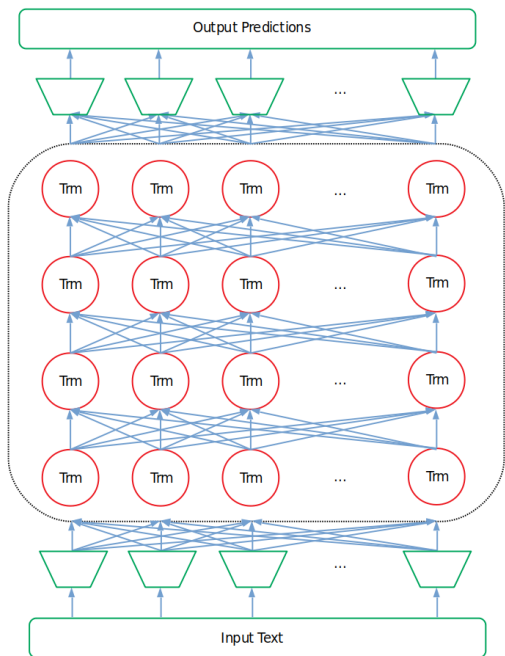
Pretrained Layer : 자연어 이해도가 높은 파라미터

Input Text : NLP Task의 Input Text

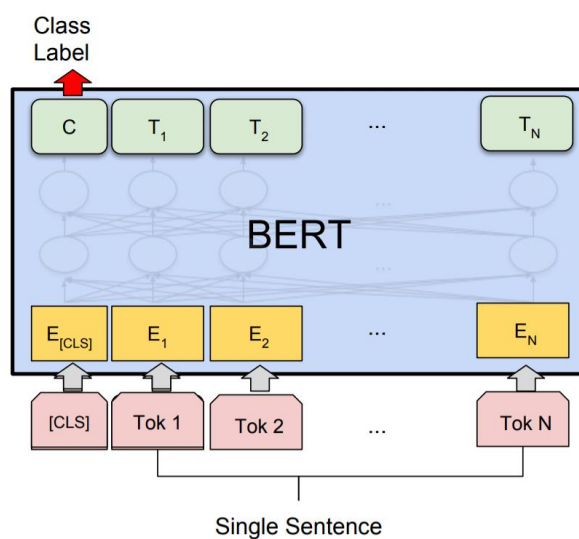


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{##ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

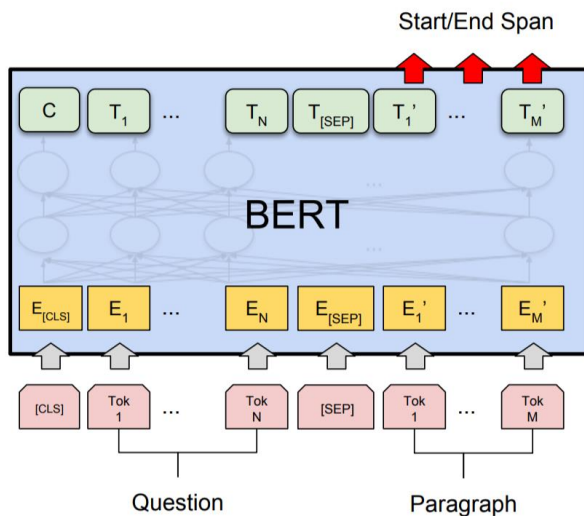
BERT 모델 개요



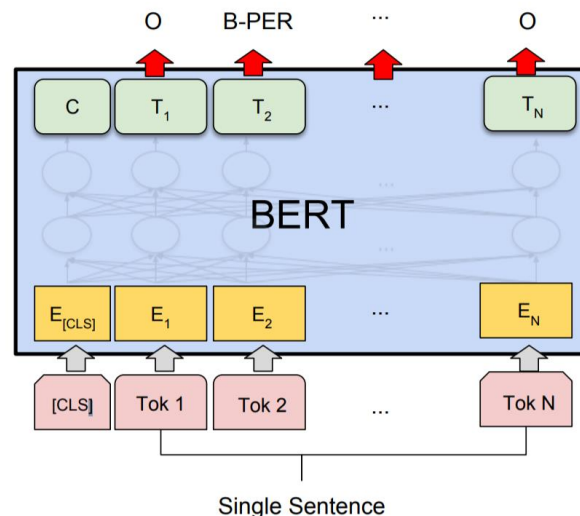
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



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



(c) Question Answering Tasks:
SQuAD v1.1



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

BERT Experiment

BERT 모델은 4가지의 실험, 총 11가지의 NLP Task에서 높은 성능을 보이며 SOTA를 달성함

GLUE (General Language Understanding Evaluation)

SQuAD v1.1 (Stanford Question Answering Dataset)

SWAG (The Situations With Adversarial Generations)

NER (Named Entity Recognition) with *CoNLL 2003 Named Entity Task*

BERT Experiment

BERT 이전의 Transformer 모델의 성능 측정

Transformer 모델의 목적은 인코더 디코더 구조를 활용한 기계번역이기 때문에 BLUE라는 기계번역의 성능지표를 내세움

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	41.29	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$3.3 \cdot 10^{18}$	
Transformer (big)	28.4	41.8	$2.3 \cdot 10^{19}$	

BERT Experiment

KorQuAD 2.0

The Korean Question Answering Dataset

Leaderboard

KorQuAD 2.0의 Test set으로 평가한 Exact Match(EM) 및 F1 score 입니다.

Rank	Reg. Date	Model	EM	F1
-	2019.09.05	Human Performance	68.82	83.86
1	2020.09.21	SDS-NET v1.3 (single model) Samsung SDS AI Research	77.86	89.82
2	2020.08.28	Ko-LongBERT (single model) LAIR	77.88	89.62
3	2020.08.28	SkERT-Large 1.1 (single model) Skelter Labs	77.44	88.81
4	2020.07.01	SDS-NET v1.2 (single model) Samsung SDS AI Research (Bae & Kwon)	76.73	88.78
5	2020.06.24	SkERT-Large (Single model) Skelter Labs	76.64	88.09
6	2020.05.03	SDS-NET v1.1 Samsung SDS AI Research (Bae & Kwon)	73.87	86.81
7	2020.04.29	SDS-NET (single model) Sanghwan Bae & Soonhwan Kwon	73.51	86.56
8	2020.08.31	anonymous-baseline (single model) Anonymous	71.24	86.37
9	2020.06.22	SkERT-Large (Single model) Skelter Labs	73.29	86.03

LGCNS AI Tech Talk for NLU (feat.KorQuAD)

<https://korquad.github.io/>

[표 5] KorQuAD 2.0 예시, □는 질문에 대한 답

예시

질문

서울의 GDP는 세계 몇 위야? (Short, 평문)

지문

서울특별시

위키백과, 우리 모두의 백과사전.

좌표: 북위 37° 34′ 00″ 동경 126° 58′ 41″

비슷한 이름의 **서울**에 관해서는 해당 문서를 참조하십시오.


서울특별시(서울特別市)는 대한민국의 수도이자 최대 규모의 도시이다. **백제**의 첫 수도인 **위례성**이었고, 고려 때는 **남경**(南京)이었으며, 조선의 수도가 된 이후로 현재까지 대한민국 정치·경제·사회·문화의 중심지 역할을 하고 있다. 중앙으로 한강이 흐르고, **북한산**, **관악산**, **도봉산**, **불암산**, **인왕산**, **청계산** 등의 여러 산들로 둘러싸인 **분지** 지형의 도시이다. 동서 간의 거리는 36.78 km, 남북 간의 거리는 30.3 km이며, 넓이는 605.25 km²이다. 면적은 대한민국 전 국토의 0.6%를 차지하지만, 약 977만 명의 인구가 살고 있어 **인구밀도**가 높다.

시청 소재지는 **종구**이며, 25개의 **자치구**로 이루어져 있다. 1986년 아시안 게임, 1988년 하계 올림픽, 2010년 **서울 G20 정상회의**를 개최한 국제적인 도시이다. 서울의 GDP는 세계 4위이다.^{[2][3]}

서울

특별시

서울特別市



질문

서울에 있는 산들에 대해 알려줘 (Long, 평문)

지문

지리

[편집]

서울은 북위 37° 34′, 동경 126° 59′의 **한반도** 중서부에 위치하는 **분지** 지형의 도시이다. 시의 중심으로 **한강**이 흐르고, 서울 도심에는 **남산**, **인왕산**(338m)이 있다. 시 주변으로 **북한산**(837m), **관악산**(629m), **도봉산**(740m), **수락산**(428m), **불암산**(510m), **구룡산**(306m), **우면산**(293m), **아차산**, **지왕산** 등이 서울을 둘러싸며 경기도 및 인천광역시와 자연적 경계를 이루고 있다.^[7] 동서 간의 거리는 36.78 km, 남북 간의 거리는 30.3km이며, 면적은 약 605.25 km²이다. 서울의 면적은 대한민국의 0.6%이며 남북한 면적의 0.265%이다. 서울특별시의 최북단은 **도봉구 도봉동**이고 최남단은 **서초구 원지동**이며 최동단은 **강동구 강일동**, 최서단은 **강서구 오곡동**이다.

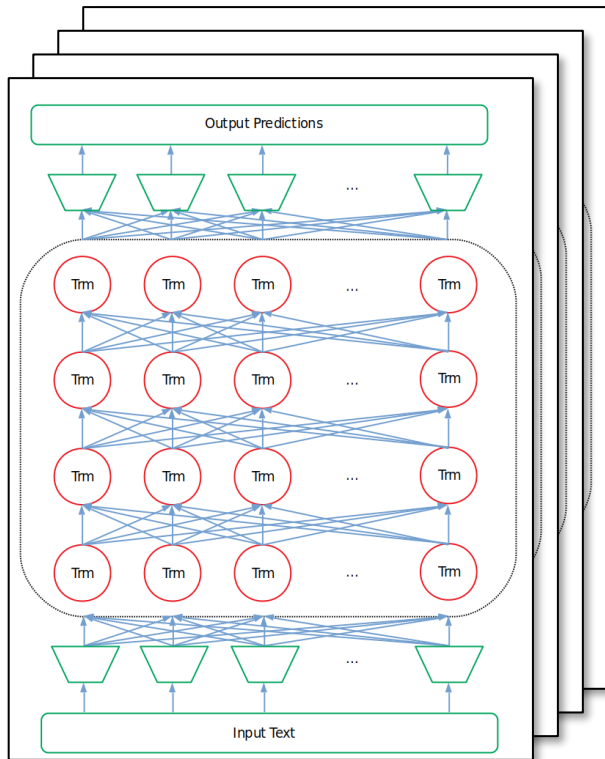
산

[편집]

한, 서울의 산, **서울** 문서를 참고하십시오.

서울은 국립공원으로 지정된 **북한산**을 최고점으로 한 고양양주구릉과 경기경야가 만나는 지대에 있다. 주위에는 **북한산**(836m)·**도봉산**(717m)·**인왕산**(338m)·**관악산**(629m) 등 500m 내외의 산과 구릉이 자연성벽과 같이 둘러싸고 있는 분지이다. **광주산맥**의 한 줄기인 **도봉산**은 백운대·인수봉·노적봉의 3개 봉우리가 솟아 있는 **북한산**과 이어져 있고, 그 산줄기는 다시 남으로 뻗어 **북악산**(342m)을 솟게 하였다. 그리고 **북악산**에서 동으로 뻗은 산줄기에 **낙산**(125m), 서로 뻗은 산줄기에 **인왕산**이 있다. **인왕산**에서 뻗은 산줄기 중 남쪽으로 뻗은 것은 **송례문**을 지나 **남산**(265m)·**충릉**(175m)과 이어져 있고, 서쪽은 무악재의 **안부**(鞍部)를 지나 **안산**(296m)과 이어져 있는데 모두 구릉성 산지이다. **한강** 남쪽에는 100m 이하의 구릉지가 펼쳐져 있고, 남쪽에 천연의 요새와 같이 서울의 외곽에 솟아 있는 **관악산**(629m)와 **청계산**(618m), **구룡산**(306m), **우면산**(293m)이 있다. 그 외에 서울 동부에 **불암산**, **수락산**, **망우산**, **아차산**이 있다.^[7]

GLUE (General Language Understanding Evaluation)



- 1 Corpus of Linguistic Acceptability (CoLA)
- 2 Stanford Sentiment Treebank (SST-2)
- 3 Microsoft Research Paraphrase Corpus (MRPC)
- 4 Quora Question Pairs (QQP)
- 5 Semantic Textual Similarity Benchmark (STS-B)
- 6 Multi-Genre NLI corpus (MNLI)
- 7 The Recognizing Textual Entailment (RTE)
- 8 The Stanford Question Answering NLI (QNLI)
- 9 The Winograd Schema Challenge NLI (WNLI)

문법 오류 발견

영화리뷰 긍정부정 파악

두 개 문장 의미 동일 파악

두 개 질문 의미 동일 파악

문장 유사도 파악

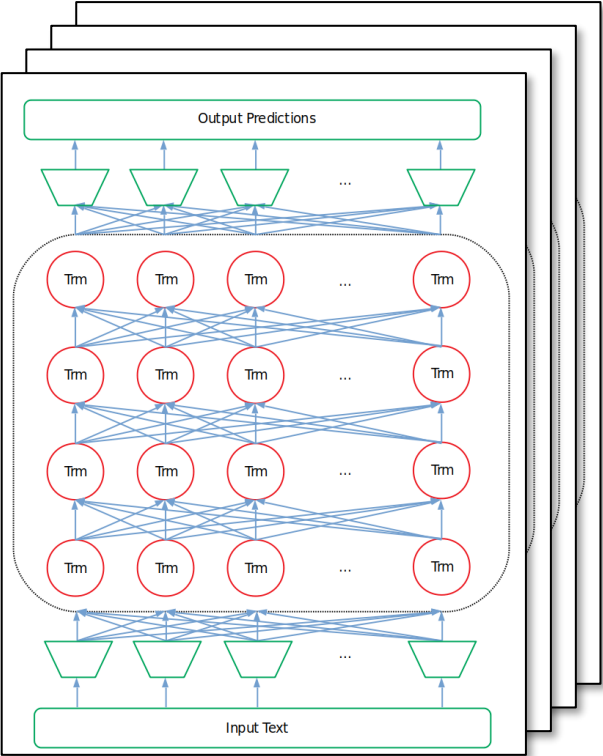
문장 간 관계 파악

문장 모순 파악















문장 모순 파악

문장 모순 파악

GLUE 데이터 예시



4	<p>Quora Question Pairs (QQP)</p> <p>input</p> <p>How do you start a bakery? How can one start bakery business?</p> <p>What are natural numbers? What is a least natural number?</p>	<p>두 개 질문 의미 동일 파악</p> <p>output</p> <p>1 (O)</p> <p>0 (X)</p>
5	<p>Semantic Textual Similarity Benchmark (STS-B)</p> <p>input</p> <p>A plane is taking off. An air plane is taking off.</p> <p>Three men are playing chess. Two men are playing chess.</p> <p>A man is smoking. A man is skating.</p>	<p>문장 유사도 파악</p> <p>output</p> <p>5.000</p> <p>2.600</p> <p>0.500</p>

Rank Name		Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX	
1	ERNIE Team - Baidu	ERNIE		90.9	74.4	97.8	93.9/91.8	93.0/92.6	75.2/90.9	91.9	91.4	97.3	92.0	95.9	51.7	
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		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	
3	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	
+	4	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	90.7	97.4	91.2	94.5	49.1
+	5	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
6	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	
7	Liangzhu ge	Deberta-xxlarge (adv + ensemble)		90.2	71.9	97.3	92.4/89.9	93.2/92.9	75.6/90.8	91.7	91.5	96.3	91.9	95.2	35.2	
8	Microsoft D365 AI & MSR AI & GATECHMT-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	
+	9	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.3	94.5	47.9
+	10	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
+	11	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
+	12	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
13	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3	
14	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7	
+	15	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
16	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-	
17	Adrian de Wynter	Bort (Alexa AI)		83.6	63.9	96.2	94.1/92.3	89.2/88.3	66.0/85.9	88.1	87.8	92.3	82.7	71.2	51.9	
+	18	Lab LV	ConvBERT base		83.2	67.8	95.7	91.4/88.3	90.4/89.7	73.0/90.0	88.3	87.4	93.2	77.9	65.1	42.9

GLUE

2018년 BERT 모델 발표 당시 모든 NLP Tasks에 대해서 SOTA를 달성함

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
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

KLUE (Korean Language Understanding Evaluation Benchmark)

인공지능 신문 (최광민 기자 2021.05.17)

<http://www.aitimes.kr/news/articleView.html?idxno=21043>

업스테이지, 한국어 기반 AI모델 공정 평가 위한 '한국어 자연어 이해 평가 데이터셋(KLUE)' 공개한다

최광민 기자 | 입력 2021.05.17 08:56 | 댓글 0

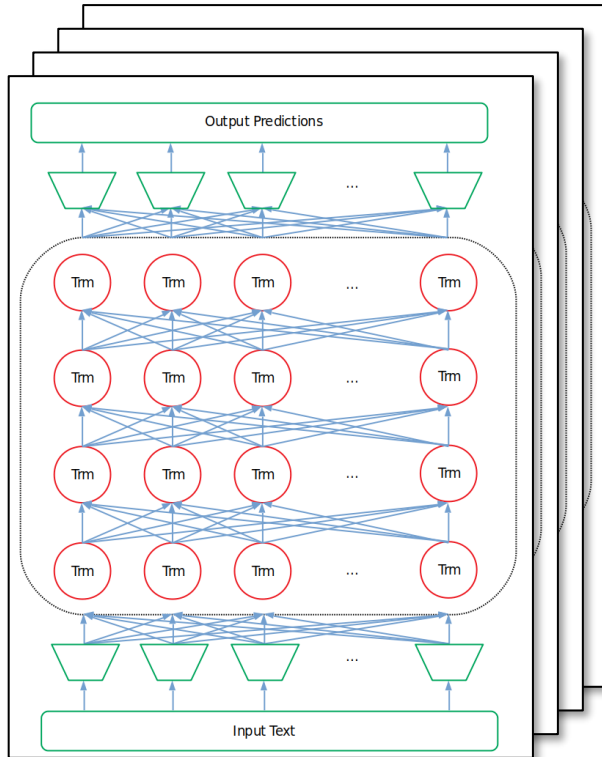


KLUE 프로젝트는 업스테이지가 호스트를 맡고, 네이버 클로바, NYU(New York University), KAIST와 공동 주최로 진행, 다양한 한국어 언어모델이 동일한 평가 선상에서 정확하게 비교될 수 있는 평가기준과 토대가 된다는 점에서 앞으로 한국어 자연어처리 분야의 발전을 앞당길 것으로 기대된다.



클루(KLUE)는 한국어 언어모델의 공정한 평가를 위한 목적으로 8개 종류의 한국어 자연어 이해 문제가 포함된 데이터 집합체이다.

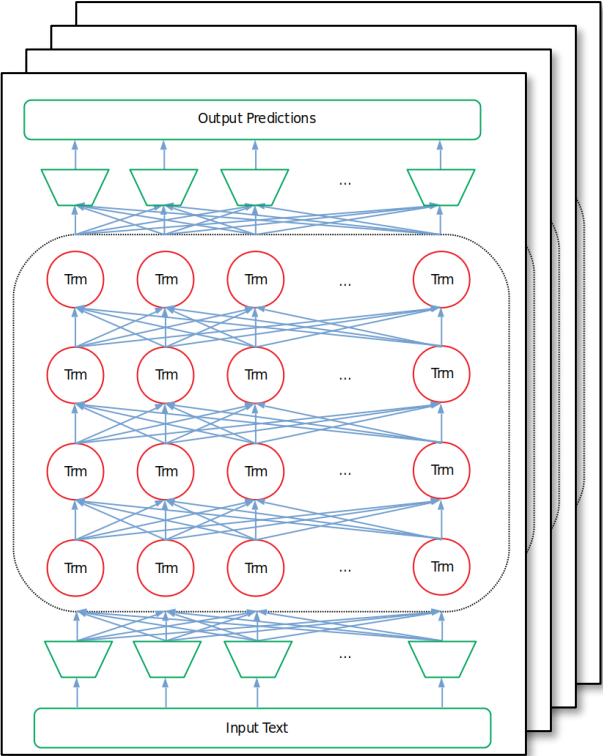
KLUE (Korean Language Understanding Evaluation Benchmark)



- 1 Topic Classification
- 2 Semantic Textual Similarity
- 3 Natural Language Inference
- 4 Named Entity Recognition
- 5 Relation Extraction
- 6 Dependency Parsing
- 7 Machine Reading Comprehension
- 8 Dialogue State Tracking

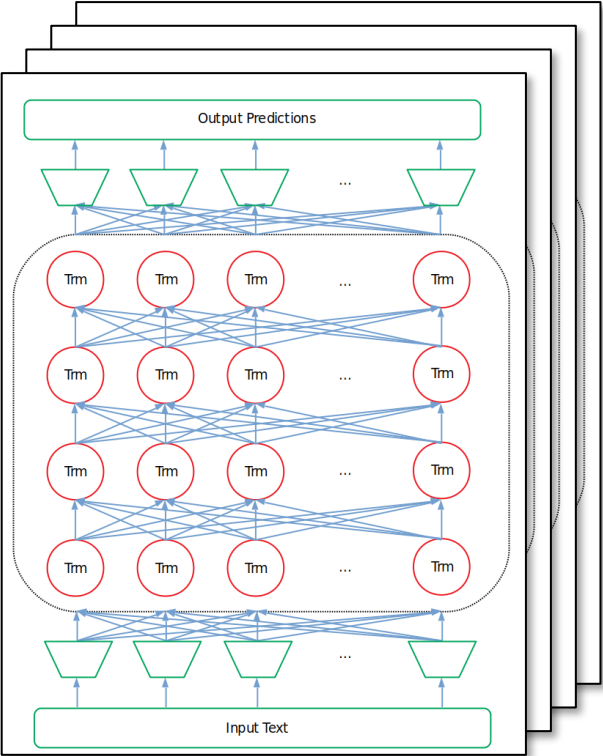
- 뉴스 헤드라인 분류
- 문장 유사도 비교
- 자연어 추론
- 개체명 인식
- 관계 추출
- 형태소 및 의존 구문 분석
- 기계독해 이해
- 대화 상대 추적

KLUE 데이터 예시



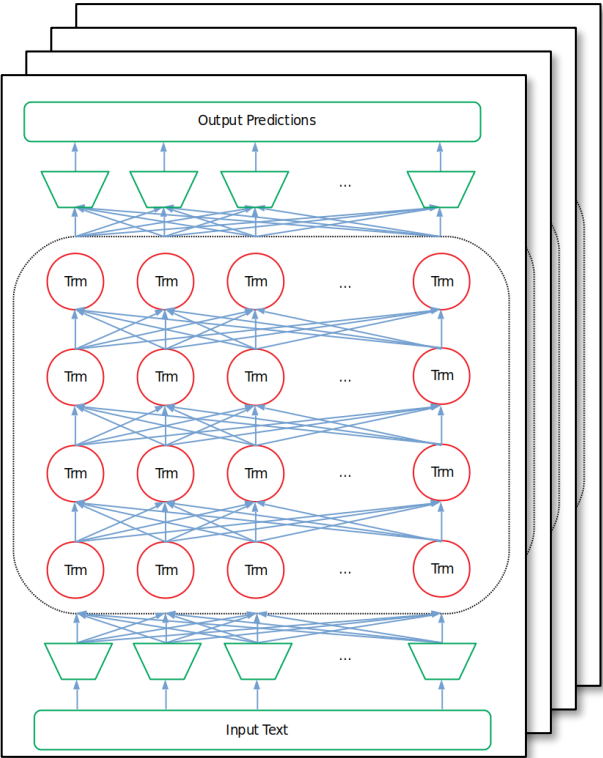
1	Topic Classification	뉴스 헤드라인 분류
	input 유튜브 내달 2일까지 크리에이터 지원 공간 운영	output 3 (Society)
2	Semantic Textual Similarity	문장 유사도 비교
	input 1: 숙소 위치는 찾기 쉽고 일반적인 한국의 반지하 숙소입니다. 2: 숙박시설의 위치는 쉽게 찾을 수 있고 한국의 대표적인 반지하 숙박시설입니다.	output 'label': 3.7
3	Natural Language Inference	자연어 추론
	input 힛걸 진심 최고다 그 어떤 히어로보다 멋지다	output 0 (긍정)

KLUE 데이터 예시



4	Named Entity Recognition	개체명 인식
	input 특히 <영동고속도로> <강릉> 방향 <문막휴게소>에서 <만종분기점>까지 <5km> 구간에는 승용차 전용 임시 갓길차로제를 운영하기로 했다.	output 12, 12, 12, 2, 3, 3, 3, 3, 3, 12, 2 ...
5	Relation Extraction	관계 추출
	input <Something>는 조지 해리슨이 쓰고 비틀즈가 1969년 앨범 《Abbey Road》에 담은 노래다.	output 'label': 0 (no_relation)
6	Dependency Parsing	형태소 및 의존 구문 분석
	input 해당 그림을 보면 디즈니 공주들이 브리트니 스피어스의 앨범이나 뮤직비디오, 화보 속 모습을 똑같이 재연했다.	output 'NNG', 'NNG+JKO', 'VV+EC', 'NNP', 'NNG+XSN+JKS' ...

KLUE 데이터 예시



7	Machine Reading Comprehension	기계독해 이해
	input Context : 올여름 장마가 17일 제주도에서 시작됐다. 서울 등 중부지방은 예년보다 사나흘 정도 늦은 이달 말께 장마가 시작될 전망이다. 17일 기상청에 따르면 제주도 남쪽 먼바다에 있는 ... Question : 북태평양 기단과 오호츠크해 기단이 만나 국내에 머무르는 기간은?	output 'answer_start': [478, 478] / 'text': ['한 달가량', '한 달']
8	Dialogue State Tracking	대화 상대 추적
	input 1: 쇼핑을 하려는데 서울 서쪽에 있을까요? 2: 오 네 거기 주소 좀 알려주세요. state : ['관광-종류-쇼핑', '관광-지역-서울 서쪽'] 3: 알려주시는김에 연락처랑 평점도 좀 알려주세요. state : ['관광-종류-쇼핑', '관광-지역-서울 서쪽', '관광-이름-노량진 수산물 도매시장']	output 1: 서울 서쪽에 쇼핑이 가능한 곳이라면 노량진 수산물 도매시장이 있습니다. state : ['관광-종류-쇼핑', '관광-지역-서울 서쪽'] 2: 노량진 수산물 도매시장의 주소는 서울 동작구 93806입니다. state : ['관광-종류-쇼핑', '관광-지역-서울 서쪽', '관광-이름-노량진 수산물 도매시장'] 3: 그럼. 연락처는 6182006591이고 평점은 4점입니다. State : ['관광-종류-쇼핑', '관광-지역-서울 서쪽', '관광-이름-노량진 수산물 도매시장']

KLUE 데이터 config

1	Topic Classification	뉴스 헤드라인 분류	train: 45,678 validation: 9,107
2	Semantic Textual Similarity	문장 유사도 비교	train: 11,668 validation: 519
3	Natural Language Inference	자연어 추론	train: 24,998 validation: 3,000
4	Named Entity Recognition	개체명 인식	train: 21,008 validation: 5,000
5	Relation Extraction	관계 추출	train: 32,470 validation: 7,765
6	Dependency Parsing	형태소 및 의존 구문 분석	train: 10,000 validation: 2,000
7	Machine Reading Comprehension	기계독해 이해	train: 17,554 validation: 5,841
8	Dialogue State Tracking	대화 상대 추적	train: 8,000 validation: 1,000

Huggingface datasets

<https://github.com/huggingface/datasets/tree/master/datasets/klue>

KLUE-benchmark datasets

<https://github.com/KLUE-benchmark/KLUE>

KLUE 데이터 예제

KLUE 벤치마크에서 파일을 열어보면 아래와 같은 json 형식으로 이루어져있음

```
{
  "guid": "klue-re-v1_dev_00035",
  "sentence": "그해에 안국동 별궁에서 열린 세자 순종과 순명효황후의 가례 때 종실의 원로로서 참석하였다.",
  "subject_entity": {
    "word": "순종",
    "start_idx": 19,
    "end_idx": 20,
    "type": "PER"
  },
  "object_entity": {
    "word": "세자",
    "start_idx": 16,
    "end_idx": 17,
    "type": "POH"
  },
  "label": "per:title",
  "source": "wikipedia"
},
```

Pororo Task Practice

No.	Task	비고
1.	Age Suitability Prediction	roberta
2.	Natural Language Inference	Roberta / brainbert / jaberta / zhberta
3.	Review Scoring	Roberta / brainbert / jaberta / zhberta
4.	Semantic Textual Similarity	Roberta / brainbert / jaberta / zhberta
5.	Sentiment Analysis	Brainbert / jaberta
6.	Zero-shot Topic Classification	Roberta / brainbert / jaberta / zhberta
7.	Dependency Parsing	posbert 구문 표지: NP(체언), VP(용언), AP(부사구), VNP(긍정 지정사구), DP(관형사구), IP(감탄사 구), X(의사 구), L(부호-왼쪽 괄호 및 따옴표), R(부호-오른쪽 괄호 및 따옴표) - 기능 표지: SBJ(주어), OBJ(목적어), MOD(관형어), AJT(부사어), CMP(보어), CNJ(접속어)
8.	Machine Reading Comprehension	brainbert
9.	Named Entity Recognition	Roberta / charbert / jaberta / zhberta - PERSON(PS), LOCATION(LC), ORGANIZATION(OG), ARTIFACT(AF), DATE(DT), TIME(TI), CIVILIZATION(CV), ANIMAL(AM), PLANT(PT), QUANTITY(QT), STUDY_FIELD(FD), THEORY(TR), EVENT(EV), MATERIAL(MT), TERM(TM)
10.	Question Generation	BART

<https://github.com/kakaobrain/pororo>

<https://kakaobrain.github.io/pororo/index.html>

<https://corpus.korean.go.kr/#down>