

| GPT-1 : Improving Language Understanding by Generative Pre-Training |

보아즈 분석 23기

김동환 이소정

2024.09.19

목 차

0
Abstract
1
Introduction

2
Related Work
3
Framework
4
Experiments

5
Analysis
6
Conclusion



언어를 이해한다는 개념이 무엇일까?

NLP Task

Textual Entailment

두 문장 사이의 논리적 관계를
판단

전제 : 모든 보아즈 구성원은 훌
륭한 사람이다

가설 : 이 발표를 듣는 사람들은
훌륭한 사람이다

Entailment, Neutral,
Contradiction 중 택 1

Question Answering

주어진 텍스트 기반, 질문에 답

문서 : 추석은 음력 8월 15일에
치르는 행사로 설날과 더불어 한
국의 주요 연휴이자 민족 최대의
명절

질문 : 추석이란?

답변 : 한국의 명절

Semantic Similarity Assessment

두 문장이 얼마나 유사한지 측정

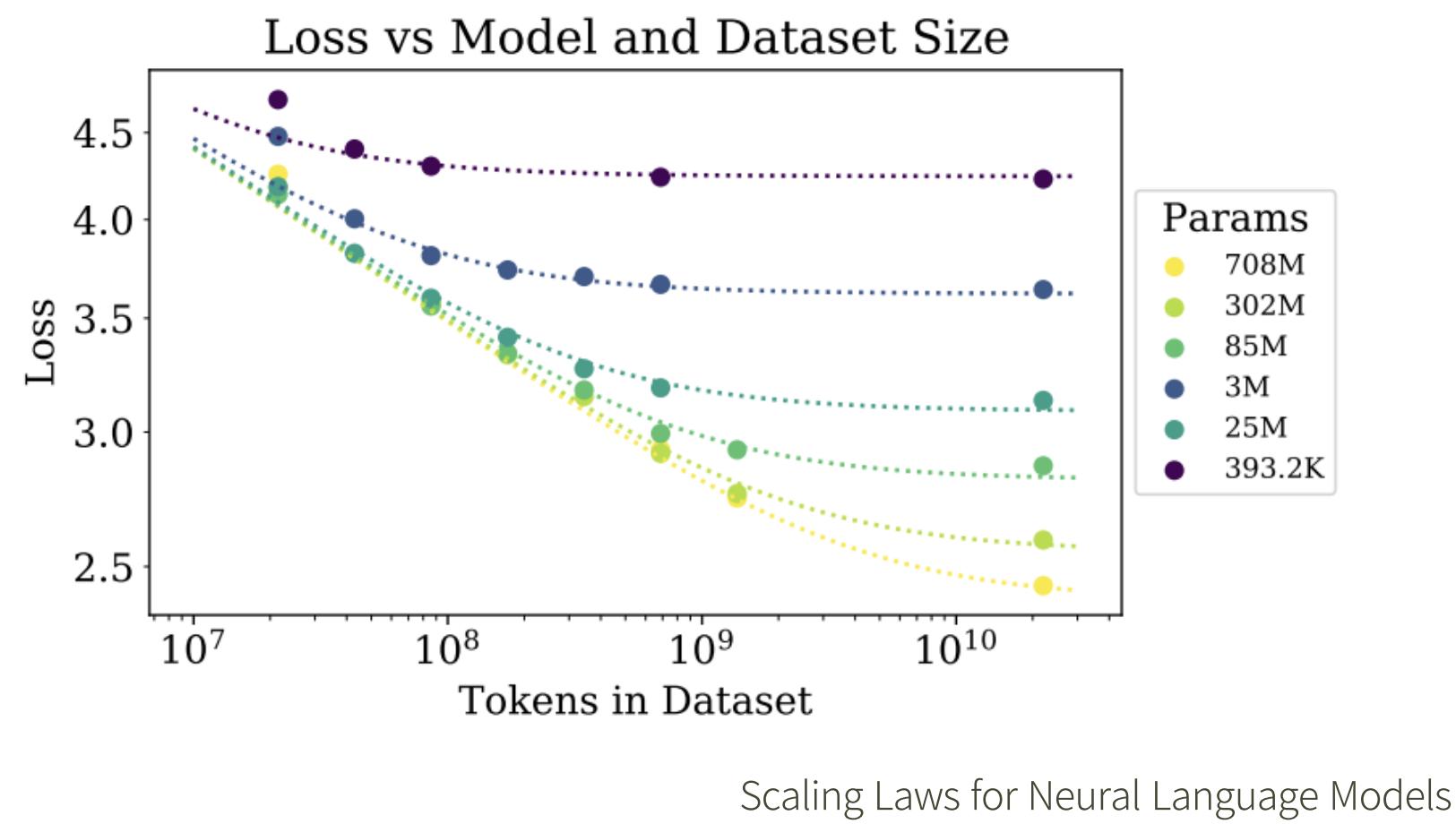
문장 1: 기차가 역에 도착했다
문장 2: 열차가 역에 도착했다

Document Classification

문서를 특정 카테고리로 분류

문서: 오늘 국제 유가가 하락하
면서 국내 주식 시장도 큰 폭으
로 떨어졌습니다

분류: 경제



데이터의 크기와 모델의 성능

Train에 사용된 Dataset의 크기가 클수록
모델의 성능이 더욱 좋아진다

문제점

라벨링 된 데이터가 별로 없음

Internet의 데이터

오답노트 고적고적 (딥러닝 기초 복습 스터디)

안녕하세요 😊 분석 23기 심현석입니다!

학기 동안 [딥러닝 기초 이론](#)을 복습하고 개념을 단단하게 가져가실 스터디원분들을 모집합니다!

▣ 스터디 모집 대상

- 이번 기회에 딥러닝의 전반적인 내용을 훑어보고 싶은 분 (분석 외 세션 활용해요!)
- 방학 중에 공부한 딥러닝 개념 및 이론들을 더 자세히 이해하고 싶은 분
- PyTorch로 CNN, RNN, LSTM 계열 모델을 구현하는 것에 관심 있는 분

▣ 스터디 내용

1. 영상: https://youtube.com/playlist?list=PLOE_1UqNACXAv9ZWMkZtv_tsvaTz5OnMe&si=nJ5twcpohBXTMuyH

- 작년에 서울대학교에서 진행된 강의 <머신러닝 및 딥러닝 1>이 유튜브에 업로드되어 있습니다. 원래 CS231n 강의로 스터디를 진행하려 하다, 한국어로 진행된 최신 강의가 있어 본 수업의 트랙을 따라가면 좋을 것 같습니다.

2. 교재: <https://www.gilbut.co.kr/book/view?bookcode=BN003345>

- 부교재로『딥러닝 파이토치 교과서』를 함께 사용할 예정입니다. 다만 스터디는 영상의 내용 위주로 진행하며, 추가적인 개념 설명이나 코드를 확인할 필요가 있는 경우 참고용으로 책을 활용하려 합니다.

실제 필요한 데이터

Passage: Raising pets is a popular online game among teenagers. ...You can feed , wash , talk to and play with your pet ...
Question: What does the passage mainly talk about?

Choices:

- A. Raising pets online is popular among teenagers .
- B. It 's bad to raise pets online .
- C. How to raise pets online .
- D. It 's good to adopt pets online .

실제 RACE 데이터셋의 일부

해결책

Generative Pre-Training

한노트 끄적끄적 (딥러닝 기초 복습 스터디)

특정 TASK에 대한 학습보다는 "일반적인" 언어 이해

- 질의응답, 함의와 같은 특정 TASK에 대해 학습시키기보다는, 보다 일반적인 언어의 **분포**를 학습시키자
 - 먼저 이렇게 학습(Pre-Training) 시킨 후, 각 작업에 대해 조금만 더 학습시키자(fine-tuning)

| GPT-1 :
Improving Language Understanding

by Generative Pre-Training |

보아즈 분석 23기
김동환 이소정

2024.09.19

GPT의 학습 과정

Step 1 : Pre-training

인터넷에 존재하는 많은 양의 Raw Data들로부터
다음 단어를 생성(Generative)하는 Task를 수행하여
언어 자체의 분포를 학습

Step 2 : Fine-tuning

사전 학습된 모델을
라벨링 된 데이터를 이용하여
특정 작업(예: 감정 분석, 문서 요약)에 맞게 업데이트

GPT의 학습 과정

Step 1 : Pre-training (Unsupervised)

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

$$h_0 = UW_e + W_p$$

$$h_l = \text{transformer_block}(h_{l-1}) \quad \forall i \in [1, n]$$

$$P(u) = \text{softmax}(h_n W_e^T)$$

Step 2 : Fine-tuning (Supervised)

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

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

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

GPT의 학습 과정

Step 1 : Pre-training (Unsupervised)

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

$$h_0 = UW_e + W_p$$

$$h_l = \text{transformer_block}(h_{l-1}) \quad \forall i \in [1, n]$$

$$P(u) = \text{softmax}(h_n W_e^T)$$

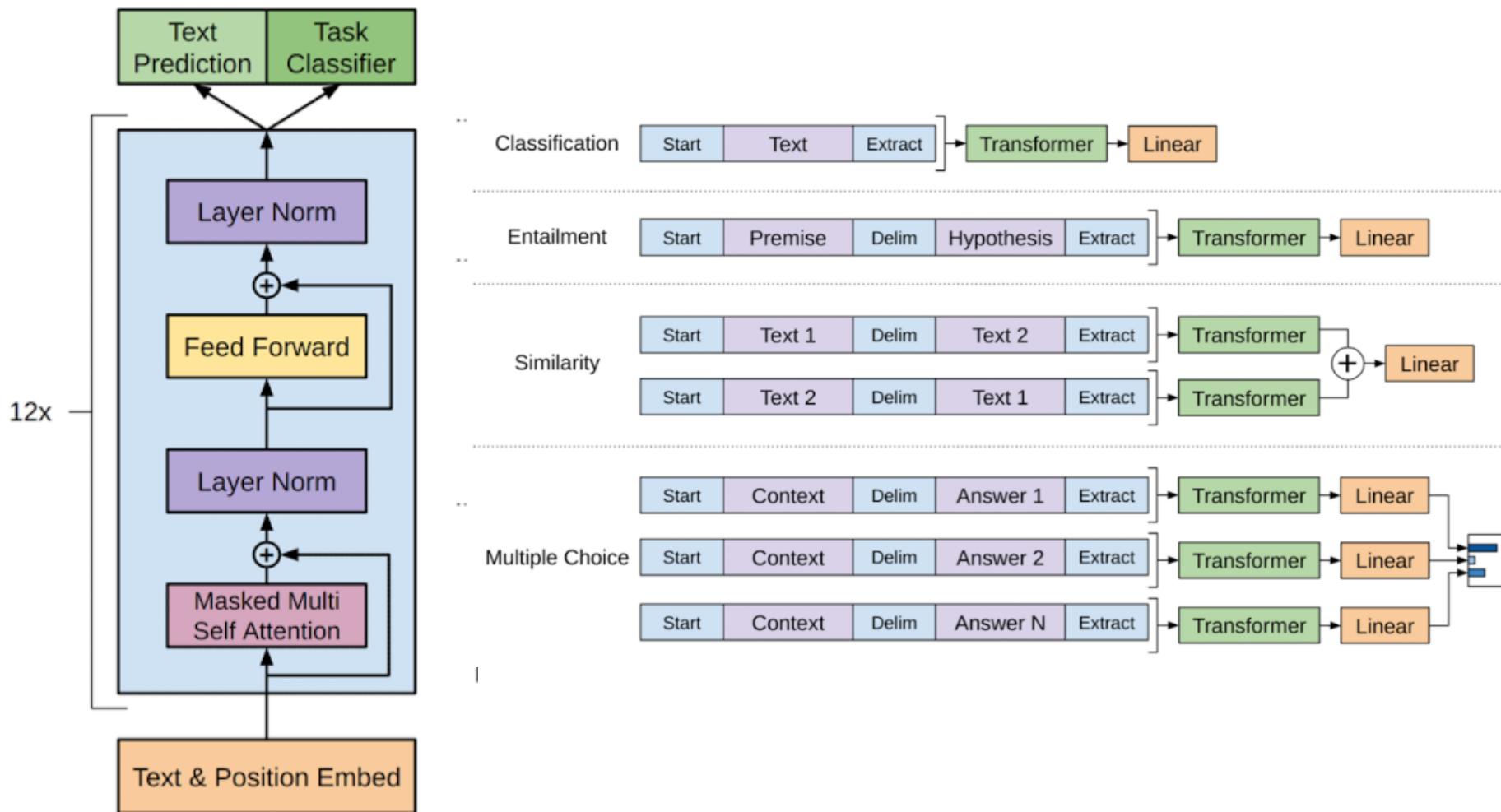
Step 2 : Fine-tuning (Supervised)

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

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

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

Fine-tuning



Fine-tuning시에는 입력 포맷이 달라짐

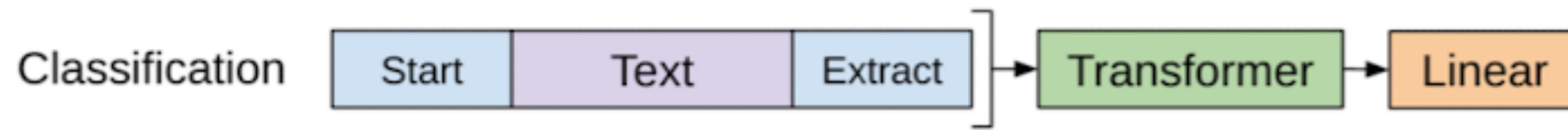
특정 Task 각각에 대해 입력을 전처리 해주어야 함

Classification



단순 Linear 계층만 추가해주면 됨

Classification

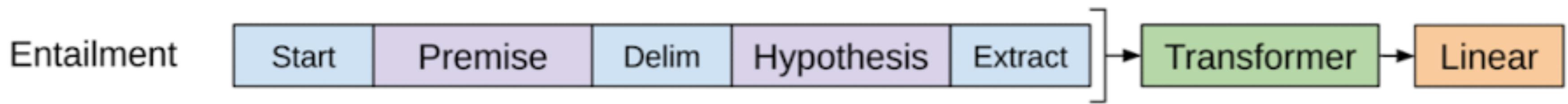


문서: "오늘 국제 유가가 하락하면서 국내 주식 시장도 큰 폭으로 떨어졌습니다."

분류: "경제"

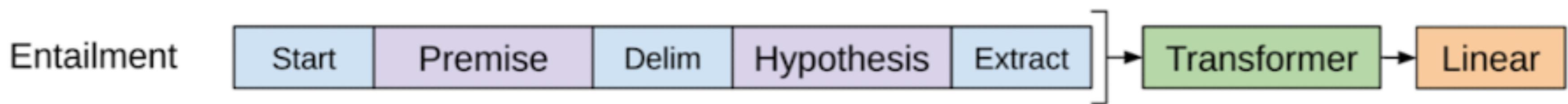
<S> 오늘 국제 유가가 하락하면서 국내 주식 시장도 큰 폭으로 떨어졌습니다. <e>

Entailment



전제와 가설을 구분자 (\$)로 구분

Entailment

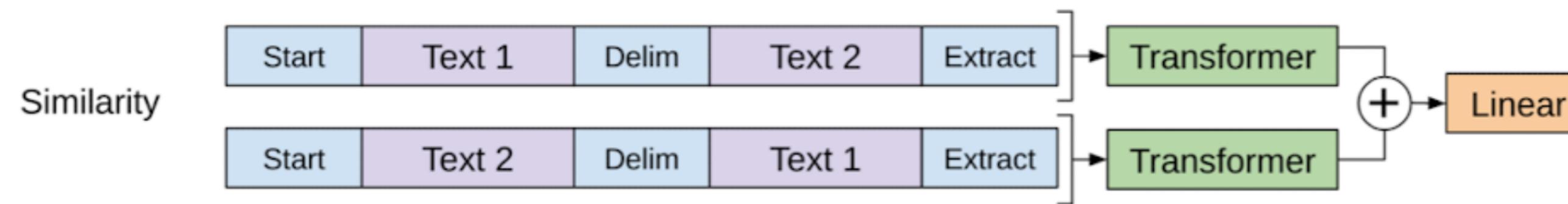


전제 : 모든 보아즈 구성원은 훌륭한 사람이다.

가설 : 이 발표를 듣는 사람들은 훌륭한 사람이다.

모든 보아즈 구성원은 훌륭한 사람이다. \$ 이 발표를 듣는 사람들은 훌륭한 사람이다.

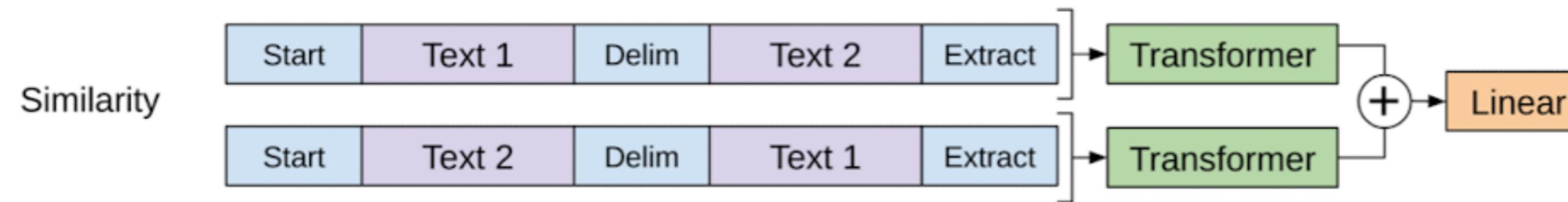
Similarity



A와 B의 유사도 == B와 A의 유사도

구분자로 구분하고, 2개 결과를 ADD

Similarity



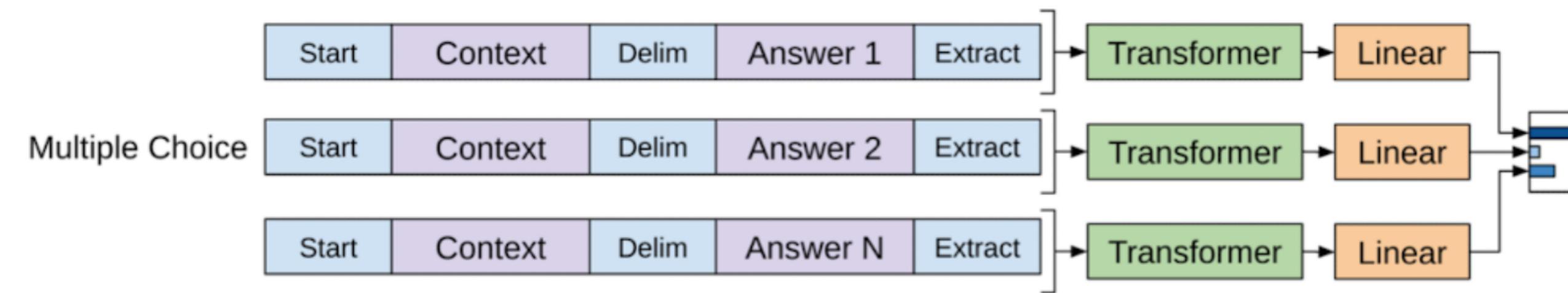
문장 1: "기차가 역에 도착했다."

문장 2: "열차가 역에 도착했다."

기차가 역에 도착했다. \$ 열차가 역에 도착했다.

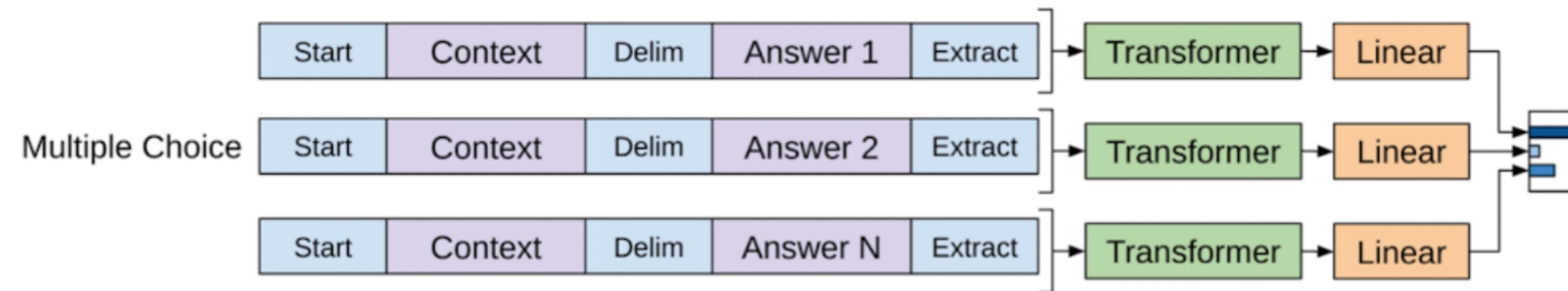
열차가 역에 도착했다. \$ 기차가 역에 도착했다.

Multiple Choice



각 질문에 대한 답변을 연결 후 Softmax

Multiple Choice

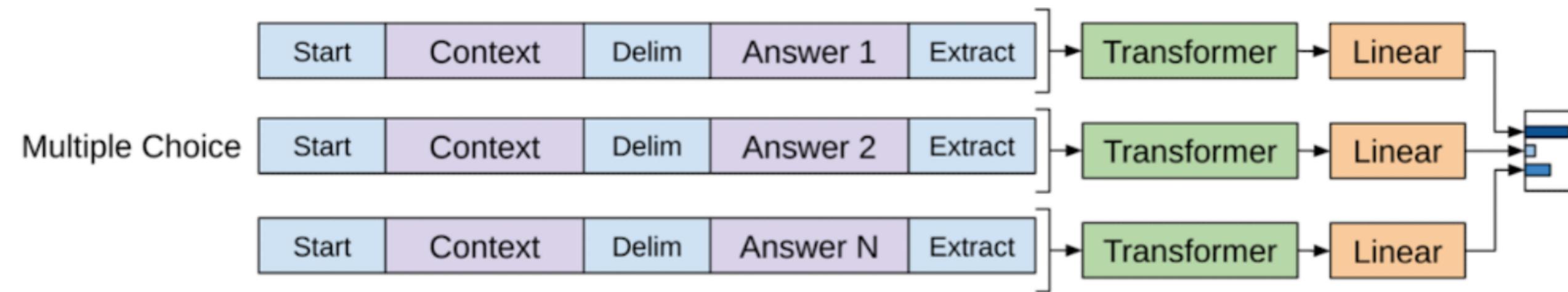


문서 : 추석은 음력 8월 15일에 치르는 행사로 한민족 최대의 명절

질문 : 추석이란?

답변 : 1) 보아즈입니다 2) 거기에 두고 가세요 3) 한국의 명절입니다

Multiple Choice



추석은 음력 8월 15일에 치르는 행사로 한민족 최대의 명절 추석이란 \$ 보아즈입니다

추석은 음력 8월 15일에 치르는 행사로 한민족 최대의 명절 추석이란 \$ 거기에 두고 가세요

추석은 음력 8월 15일에 치르는 행사로 한민족 최대의 명절 추석이란 \$ 한국의 명절입니다

| GPT-1 : Experiments |

보아즈 분석 23기
김동환 이소정

2024.09.19

Experiments

unsupervised pre-training

- BookCorpus dataset which allows the generative model to learn to condition on long-range information.
 - 즉, long term dependency 학습 가능한 긴 지문 데이터셋)

Model specification

- transformer decoder layer는 총 12층
- self-attention head는 총 12개의 heads로 구성
- position-wise feed-forward는 총 3072 차원이며 adam optimizer를 사용했다.
- fine tuning은 λ 를 0.5로 설정한 것을 제외하고는 나머지 하이퍼파라미터는 unsupervised와 거의 동일하다.

Experiments

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

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

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	-	-	<u>82.1</u>	61.7
Finetuned Transformer LM (ours)	82.1	81.4	89.9	88.3	88.1	56.0

task 12개 중에 9개의 dataset에서 SOTA를 달성

natural language inference : 문맥적인 함의를 읽어내는 것, 문장 간 관계를 알아내는 것 (전제, 반대, 중립)은 다양한 단어의 뜻, 상호 참조(누가 누구를 가리키는 알아내는 것), 단어와 구문의 모호함 때문에 어려움

→ 많은 문장과 언어적 모호함에도 불구하고 gpt-1은 sota 달성!

question answering and commonsense reasoning : single and multi-sentence reasoning 이 필요함.
RACE dataset(중고등학교 시험용 영어 문단과 퀴즈 자료) + story cloze test (여러 문단 이야기의 문장 끝 완성하는 test) 를 이용하여 sota 달성!

Experiments

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

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

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	-	-	<u>82.1</u>	61.7
Finetuned Transformer LM (ours)	82.1	81.4	89.9	88.3	88.1	56.0

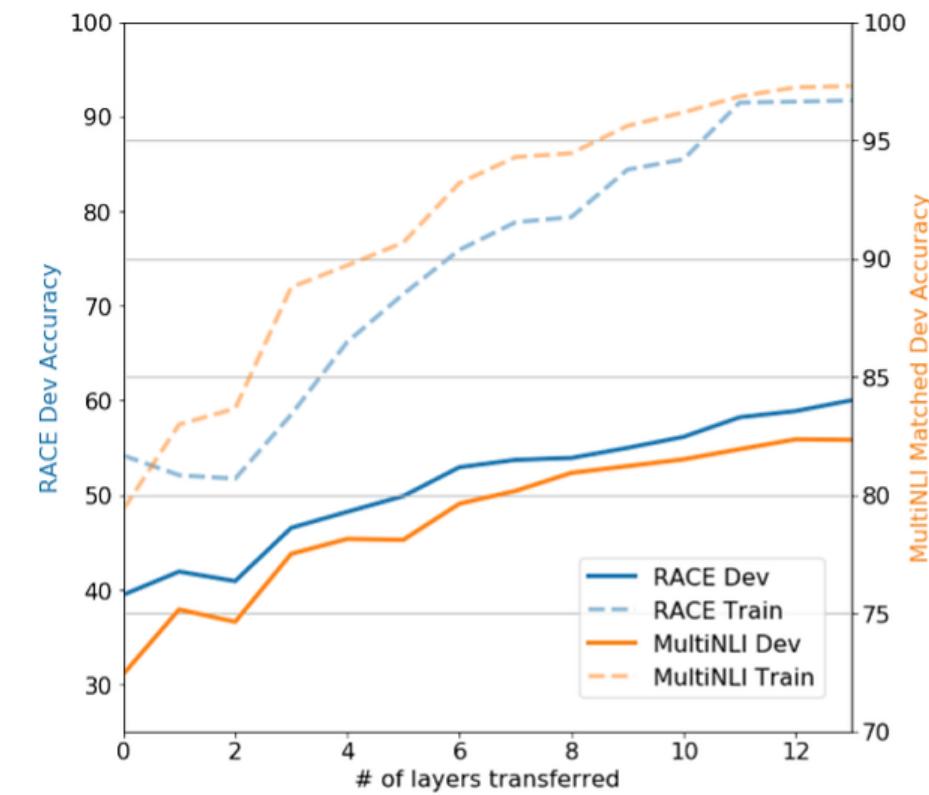
task 12개 중에 9개의 dataset에서 SOTA를 달성

semantic similarity : 두 문장이 구문론적으로 일치하나? 를 판단
개념이 rephrasing 되었는지 판단, 부정을 이해하는 것, 구문론적
모호함을 다루는것이 관건

classification : the corpus of linguistic acceptability 는 이 문
장이 문법적으로 옳은지 아닌지를 판단한 것을 담고 있으며 이를
이용해 pre-trained 된 모델이 문법적으로 편향이 있는지를 테스
트한다. 정확도가 아주 높게 나왔다 (sota)

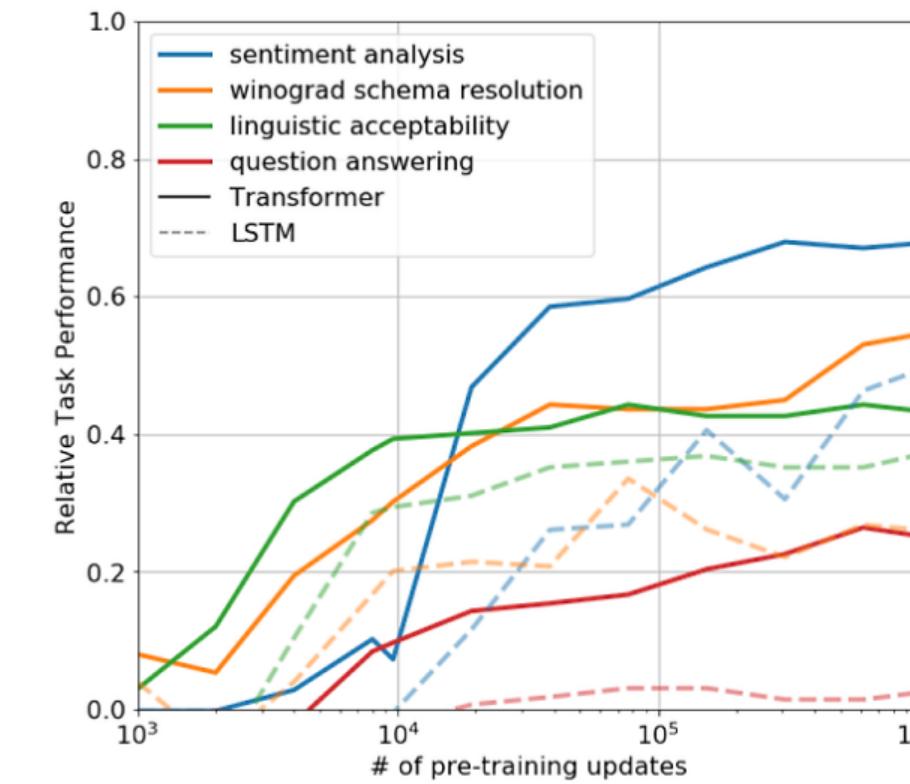
Analysis

layers수의 증가량에 따라 RACE와 MultiNLI에 대해 실험



pre-trained 된 GPT의 레이어를 몇개를 가져올까?
layer의 개수가 증가함에 따라 정확도가 향상된다.
layer # 120이후부터는 수렴 양상을 보인다.

transformer vs LSTM



LSTM을 pre-train 시킨 것보다 GPT-1(transformer)이 높은 성능을 보여주고 있다.

Analysis

Sentiment analysis

문장에 very 토큰을 append 시키고 출력을 positive, negative로 제한하였다.

Winograd schema resolution

명확한 대명사 2가지를 가능한 참조로 바꾸고 생성 모델이 치환 후 나머지 시퀀스에 대하여 더 높은 토큰의 로그 확률을 할당한다.

Linguistic Acceptability

출력 토큰의 average token log-probability를 이용하여 thresholding으로 예측

Question answering

문서와 질문을 입력으로 하여 생성되는 문장들 중에 가장 average log-probability가 높은 문장을 예측한다.

Analysis

Model performance on different tasks

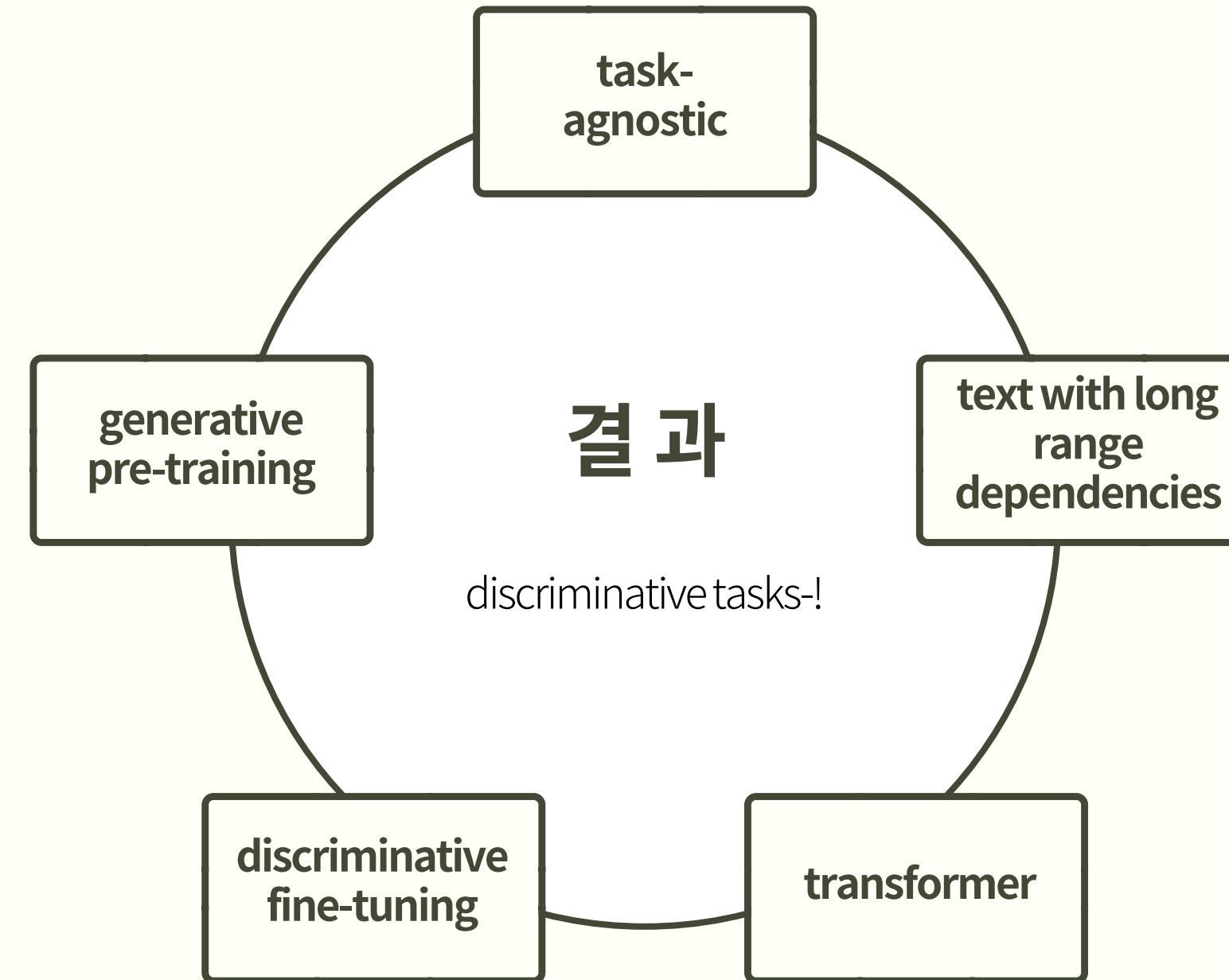
Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	70.3	81.8	88.1	56.0
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	75.0	47.9	92.0	84.9	83.2	69.8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

Datasets

- CoLA(linguistic acceptability), SST2 : classification
- MRPC, STSB, QQP : semantic similarity
- MNLI, QNLI : natural language reference task data
 - 언어 모델이 언어 이해나 생성 작업에서 얼마나 좋은 성능을 보이는지 평가하기 위한 기준이 되는 데이터

- pretraining 과정이 없을 때 성능이 15% 하락
- fine-tuning : 데이터 셋이 클 경우에 더 효과적 (1번째는 fine-tuning 한 것, 3번째 열은 안한 것)
 - 끝에 갈수록 데이터셋이 더 큼 (특히 4개)

Conclusion



감사합니다

Thank you