Improving Language Understanding by Generative Pre-Training

Name
조병웅

NLP

2024/05/14

TRAIN AND TEST

Contents

TRAIN AND TEST

Abstract

NLP에는 다양한 Task 존재! 그러나,

Labeled Data는 많이 없음.

- Text Classification
- Text Entailment
- Similarity
- Multiple Choices

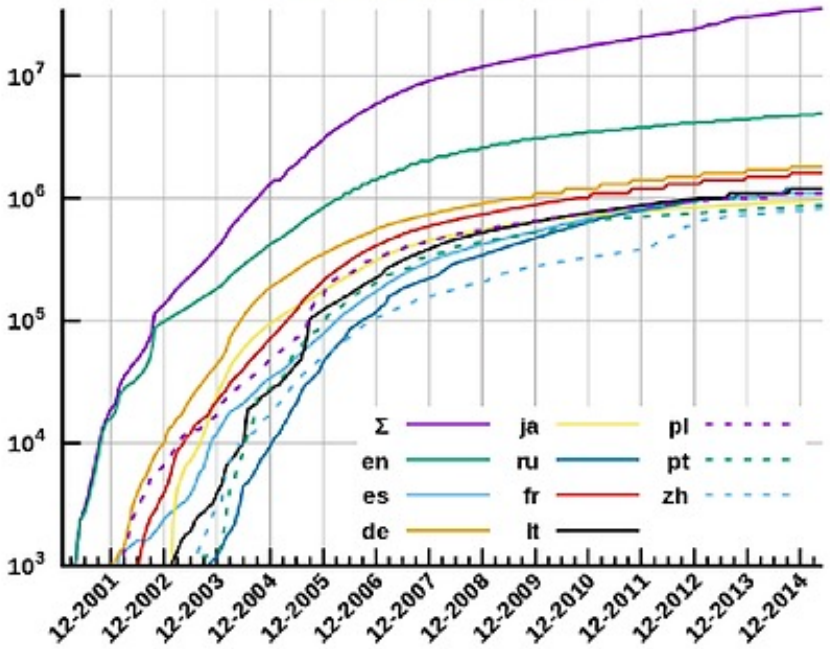
Solution -> 세상에 넘치는 Unlabeled Data를 활용해 LM 구축해보자!

TRAIN AND TEST

Abstract

Unlabeled vs Labeled

Unlabeled dataset



As of 24 February 2020, there are 6,020,081 articles in the English Wikipedia containing over 3.5 billion words.

Labeled dataset

- STS Benchmark for sentence similarity: 8,628 sentences
- Quora question pairs: 404,290 question pairs
- CoLA dataset: 10,657 sentences

TRAIN AND TEST

Abstract

GPT is Generative Pre-Training

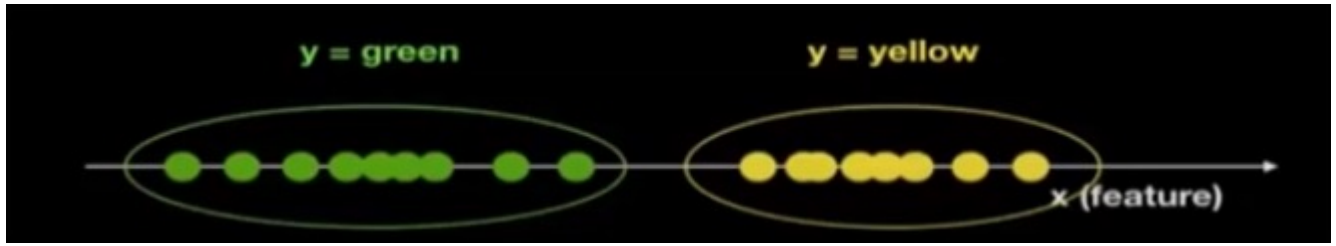
Generative Learning vs Discriminative Learning

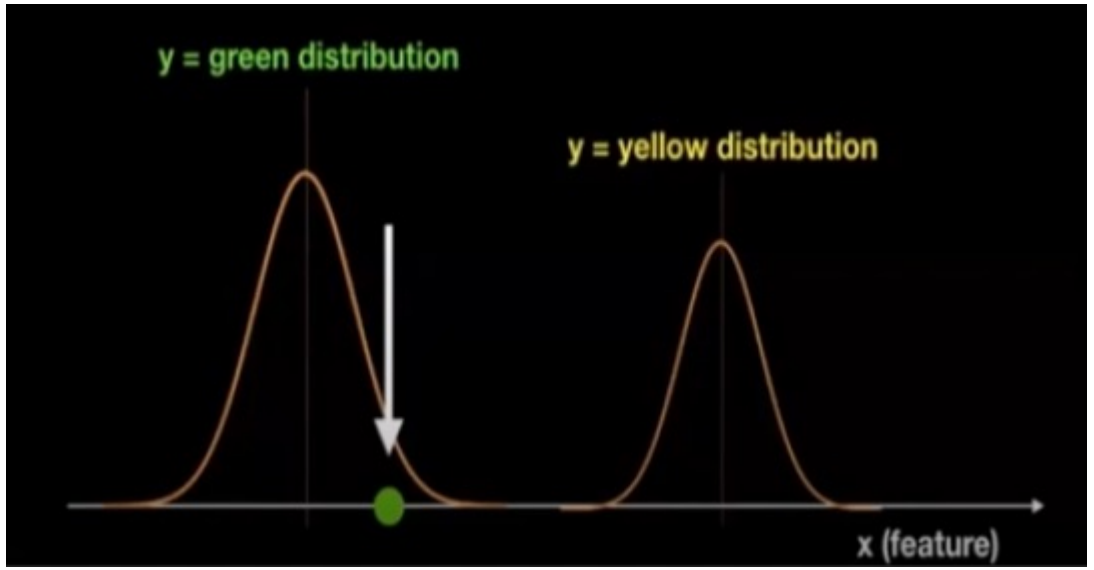
What is Pre-Training?

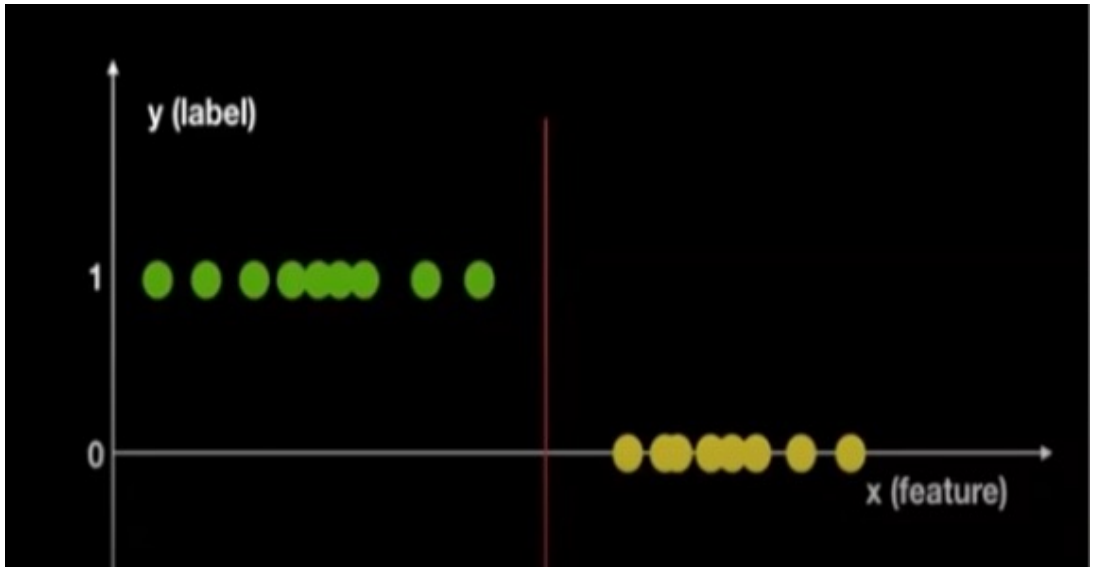
TRAIN AND TEST

Abstract

Generative Learning vs Discriminative Learning







TRAIN AND TEST

Abstract

Generative Learning vs Discriminative Learning

Generative Learning

라벨값의 분포를 활용, 데이터 '분포'를 모델링
데이터가 더 많을 때 좋고, 데이터 과적합 확률이 적음
So, 실제 분포를 나타낼 만큼 충분한 데이터, 충분한 훈련 시간 필요.

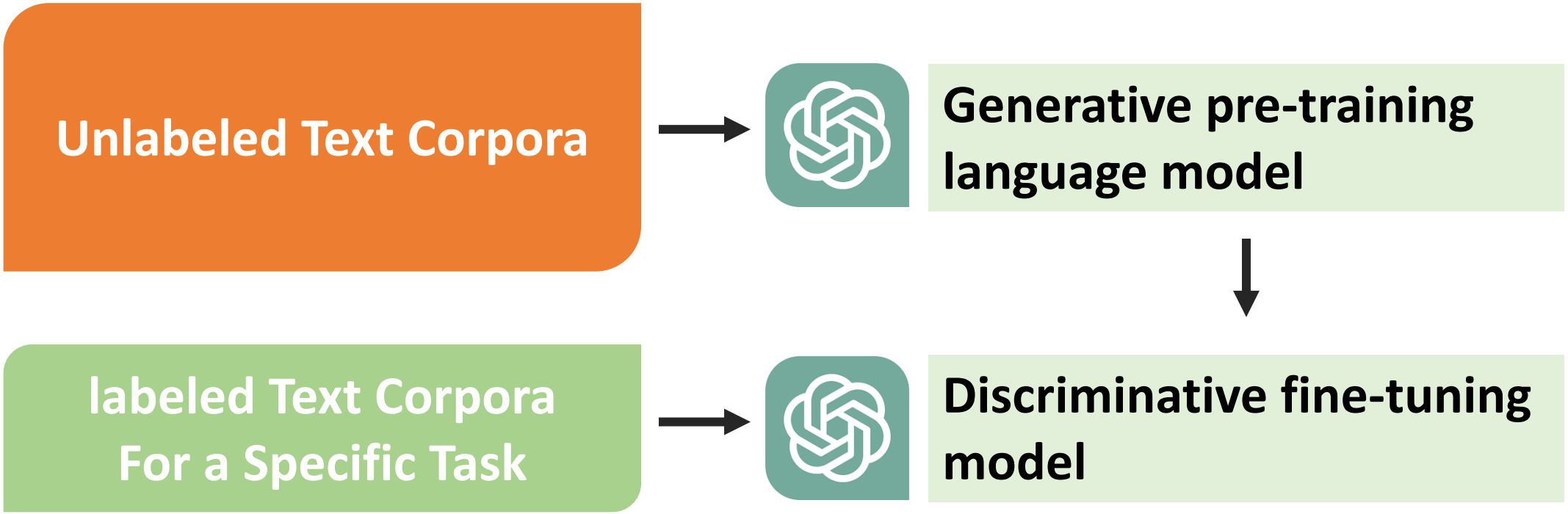
Discriminative Learning

라벨값에 따른 데이터 간의 decision boundary를 정의
데이터가 적어도 학습 가능
샘플에 대한 과적합 확률이 높음

TRAIN AND TEST

Abstract

결국 GPT는..



"Semi-Supervised Learning"

Introduction

Why Semi-Supervised Learning?

➔ **Leveraging more than word-level information from unlabeled text is challenging**

- **Transfer에 유용한 text 표현을 배우는 것에 어떤 형태의 최적화 목적(Optimization Objective)가 가장 좋은 지 불분명함 (정보부족, 목표의 불확실성, 부적절한 특성 등)**

- **학습된 표현들을 목표 작업(target task)으로 transfer하는 가장 효과적인 방법이 불분명함.**

이를 해결하고자 기존 연구에서는 Glove, discource coherence 그리고 task-specific change를 적용하는 방식 등이 제안됨 -> 이번 논문에서는 semi-supervised를 그 해결책으로 제시.

TRAIN AND TEST

Introduction

Why Semi-Supervised Learning?

➔ 최종적인 목표는 다양한 task에 작은 변화만으로 적용할 수 있는 representaion을 학습하는 것.

과정과 의미

- K size context window 설정
- 특정한 i번째 단어를 예측하기 위해 i-1부터 i-k까지 단어를 보고, 그 가능성을 최대하는 방법
  (즉, 라벨이 없는 데이터에도 학습할 수 있도록 비지도 학습을 설계)
- 확률을 최대화하는 것이기 때문에 MLE(우도 최대화) 기법을 loss function으로 설정
- SGD를 이용해 역전파

Introduction

Why Transformer(decoder)?

이유

- RNN과 비교할 때, 멀리 떨어진 요소들 사이의 의존성을 학습하기에 좋음
- 전이 학습 시에 각 작업에 맞게 입력 데이터를 변환하는 방법을 사용하여 모델을 조정함.
- Encoder는 벡터로 출력되는 반면, Decoder는 확률값으로 표현되기에 예측이 가능.
- 답변, 응답, 해석, 번역 등의 분야에서 뛰어난 구조임.

정리하면, 비지도 학습을 통해 대용량의 text 데이터를 Decoder 에 학습시켜서 LM으로 사전 학습한 것이 바로 GPT.

TRAIN AND TEST

Framework

학습은 두 가지 단계(STAGE)로 진행됨.

1. 대량의 말뭉치로 대용량의 언어 모델을 사전 학습
2. Labeled Data를 사용하여 목표 작업에 맞게 미세조정

TRAIN AND TEST

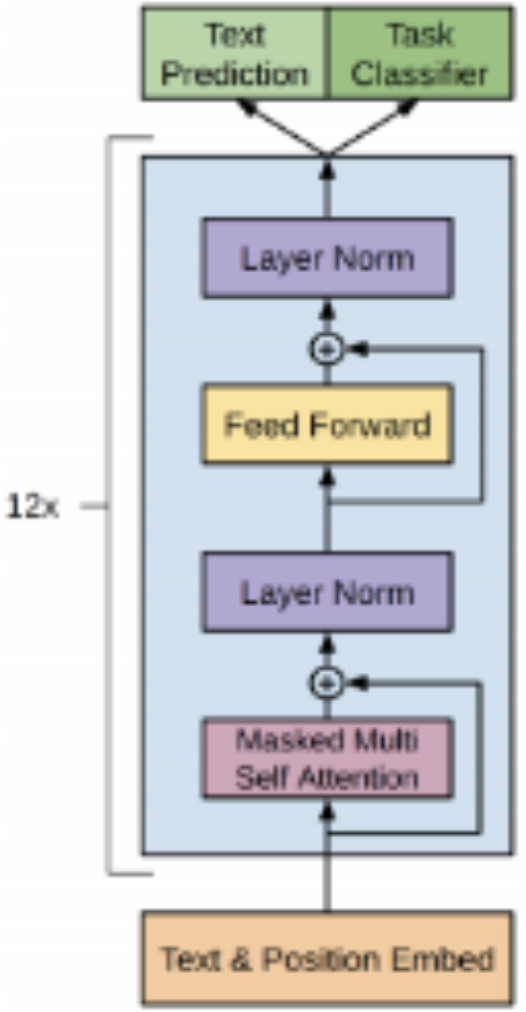
Framework

Unsupervised pre-training

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

1. input sequenc를 받아 word embedding, positional embedding 수행
   -> 이때, h_0이 위와 같이 표현됨
2. 그 다음 hidden state를 decoder 블락에 계속 넣어 학습시킴
3. 마지막 최종 hidden state 값을 활용하여 확률값 출력

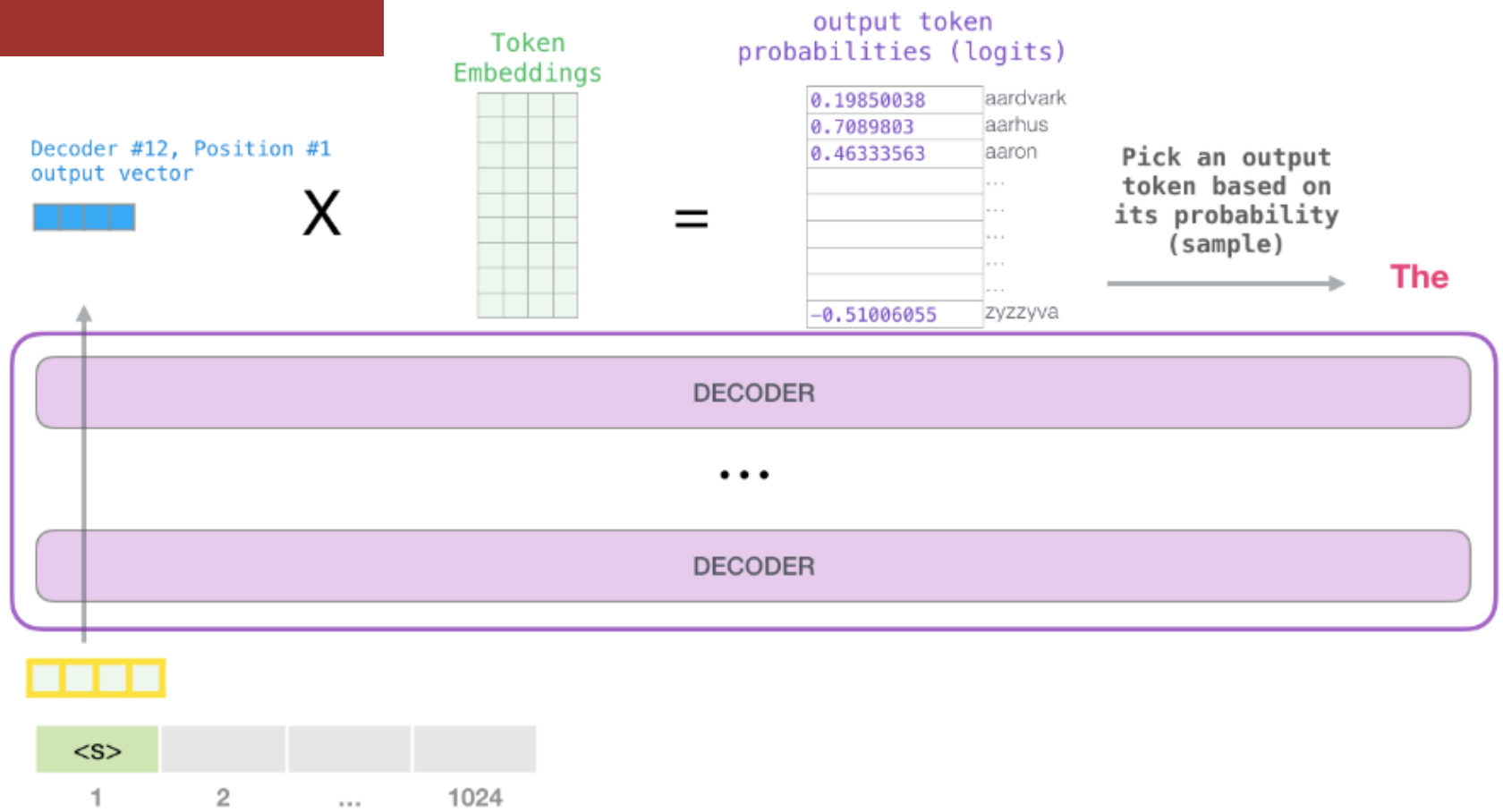
* U는 context vector, We는 토큰 임베딩 matrix, Wp는 positional matrix
* No multihead attention, just masked multihead attention.



TRAIN AND TEST

Framework

Unsupervised pre-training



Framework

Supervised fine-training

- 라벨 데이터를 활용하여, Task에 맞게 model을 fine tuning시키는 과정

the final transformer block's activation h_l^m, which is then fed into an added linear output layer with parameters W_y to predict y:

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

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

* 데이터 셋에 label y가 존재하기 때문에, y에 대한 확률을 사용.
* C는 라벨링된 데이터 셋을 의미

TRAIN AND TEST

Framework

Supervised fine-training

- 미세조정의 학습을 돕는 보조 목적 함수

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

*L2 Fuction에 L1(C)를 추가적으로 사용할 때, 일반화와 학습속도 향상에 도움.

TRAIN AND TEST

Framework

Task-specific input transformation

- 특정 task의 경우, 구조화된 입력을 제공해야 함.
- Task Specific하게 Tuning위해, Input transformation 진행
- linear+softmax의 layer만 추가하고 그 이전은 freeze.

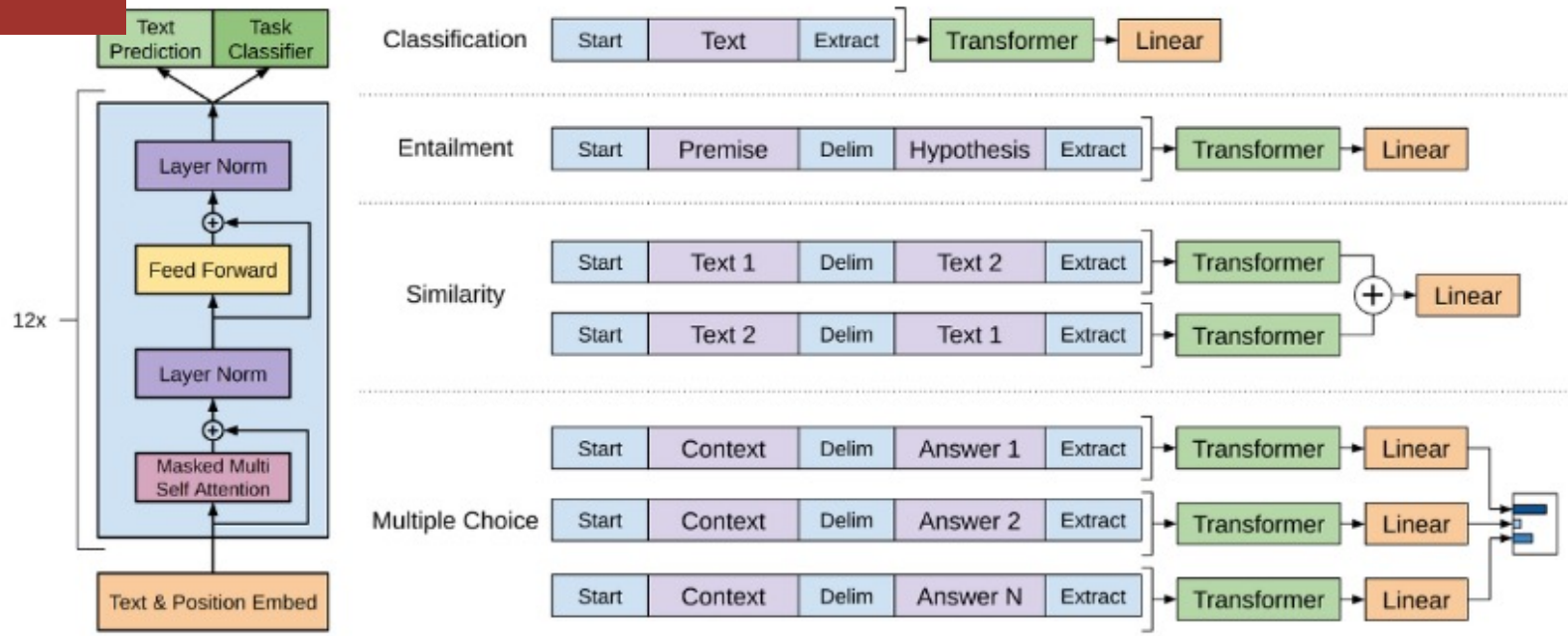


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

TRAIN AND TEST

Experiments

Setup

- 7000개의 다양한 종류의 서적 데이터 – BookCorpus Dataset
-> 상대적으로 긴 길이의 연속적 텍스트를 포함하여 long – range info를 습득

- 12개의 Transcormer decoder block + adam optimizer
-> 전반적으로 Transformer의 구조를 따름.

- Bytepair Encoding(BPE)사용
-> 데이터의 효율적 표현 가능

TRAIN AND TEST

Supervised fine-tuning

Natural Language Inference(NLI)

-> 두 문장 간의 관계를 파악하는 Task

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	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

[표-2]

TRAIN AND TEST

Supervised fine-tuning

Q n A and Commonsense reasoning

-> Multiple Choice Task

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

[표-3]

TRAIN AND TEST

Supervised fine-tuning

Classification and Semantic similarity

-> 분류와 문장간의 유사도 측정 task

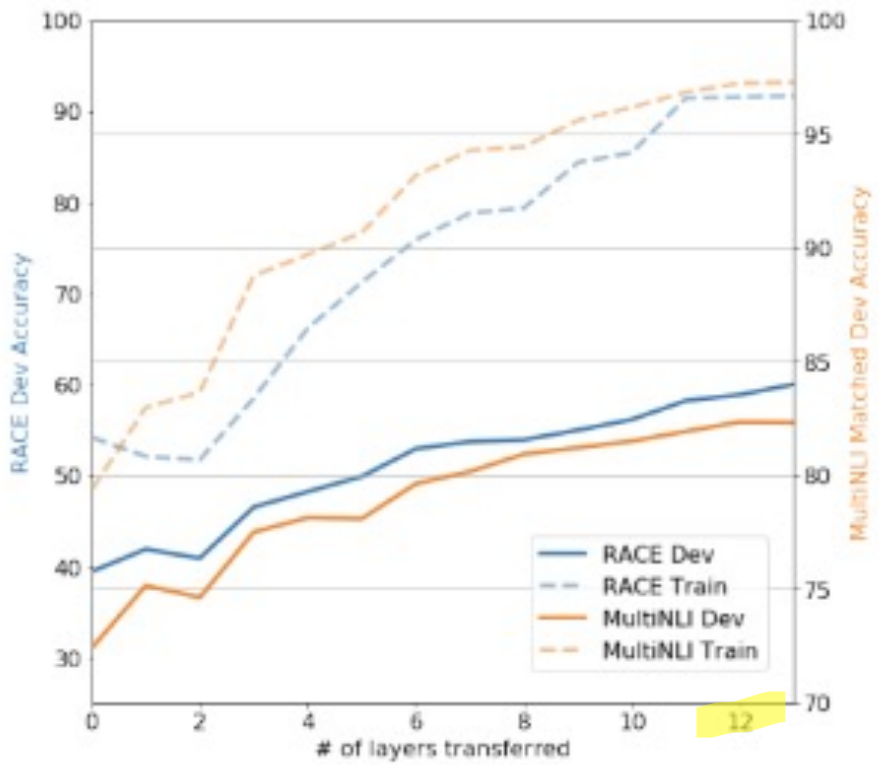
Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM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

[표-4]

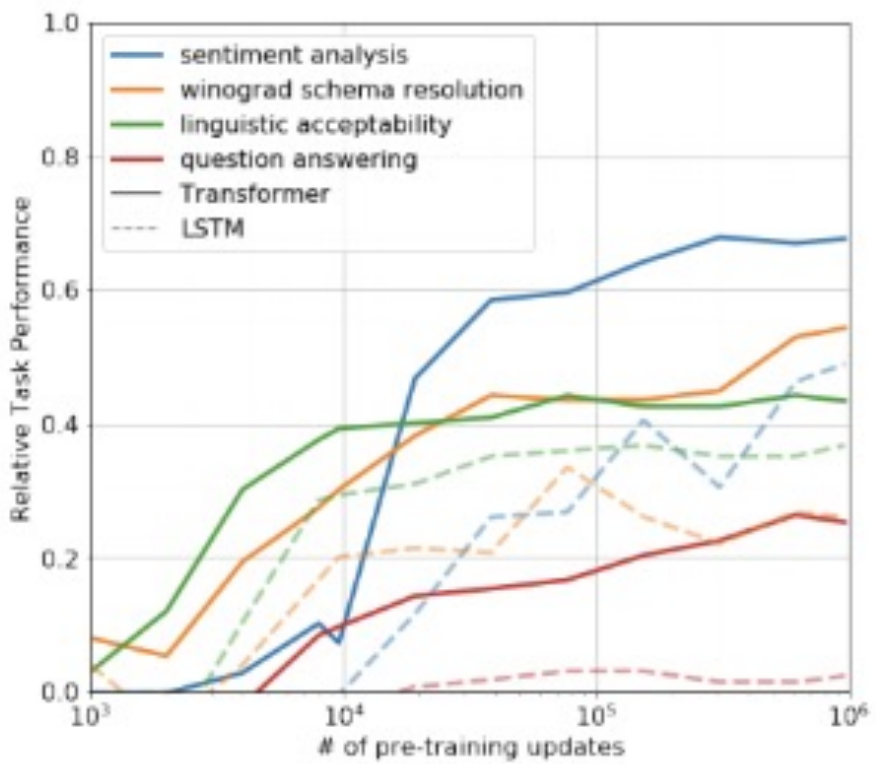
TRAIN AND TEST

Analysis

Impact of number of layers & Zero shot Behaviors



* 디코더 블락을 쌓을 수록 성능이 올라감



* 사전훈련할수록 성능이 올라감

TRAIN AND TEST

Conclusion

- GPT는 semi-supervised learnin을 사용하여 task specifi한 LM을 구축한 것.
- 광대한 정보를 담은 연속된 텍스트로 이루어진 다양한 말뭉치들로 사전학습된 모델은 상당한 일반지식과 넓은 범위의 정보 처리 가능한 기능을 얻음
- 모델을 구성할 때, unlabeled data를 활용하여 unsupervised pre-training이 가능함을 보여줌.
- 다양한 벤치마크에서 SOTA를 보여줌.

TRAIN AND TEST

TRAIN AND TEST