GPT 1 :
Generative Pre-Training

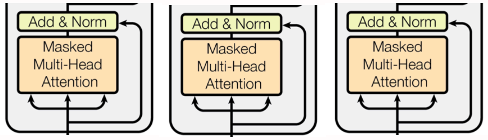
by OpenAI

NLP2조

김유진, 문예진, 송경민, 이상민, 한유경

Generative Pre-Training Language Model

unlabeled Text → pre-training → GPT LM



labeled Text
(Specific task) → fine-tuning → task

target 예측

GPT Unsupervised Pre-Training

unlabeled Text

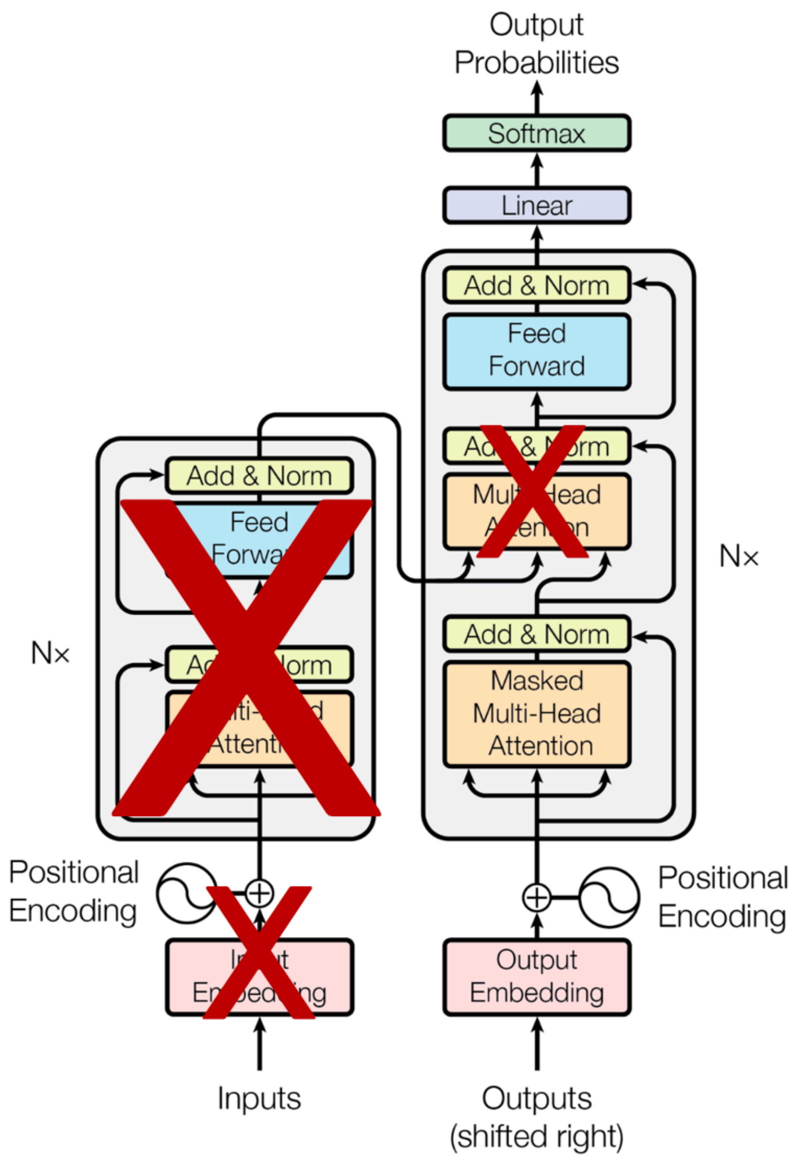
$$\mathcal{U} = (u_1, u_2, \ldots, u_n)$$

일반적인 standard LM의 목적함수 $L_1(\mathcal{U})$ likelihood 최대화

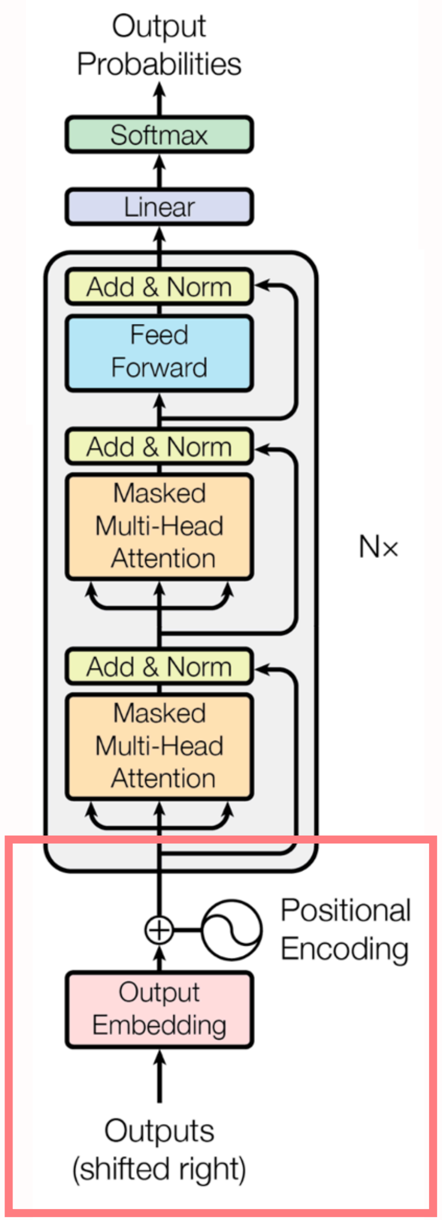
$$L_1(\mathcal{U}) = \sum_i logP(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

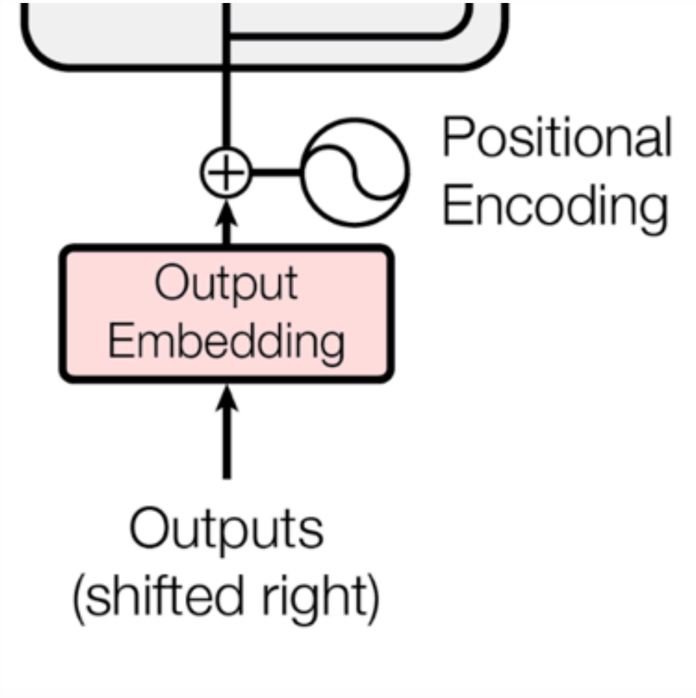
k = 윈도우 크기, 어디까지 살펴보는지

P = 파라미터 Θ를 가진 NN에 대한 조건부 확률



GPT Unsupervised Pre-Training





$$h_0 = UW_e + W_p$$

U = token의 context vector

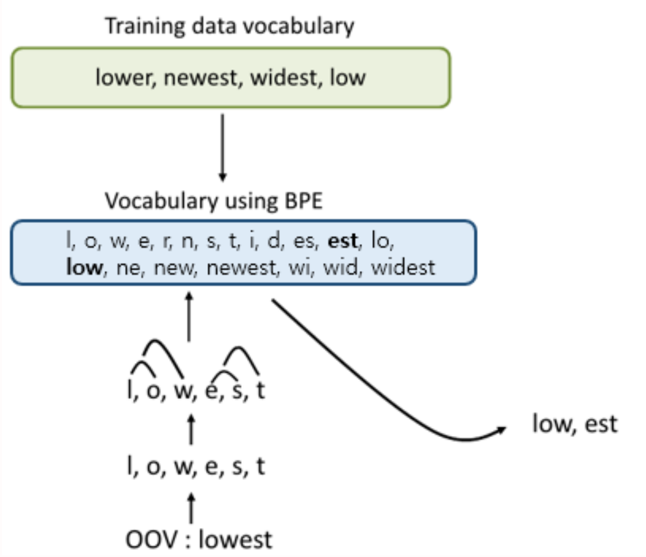
n = layer 수

W_e = token embedding matrix

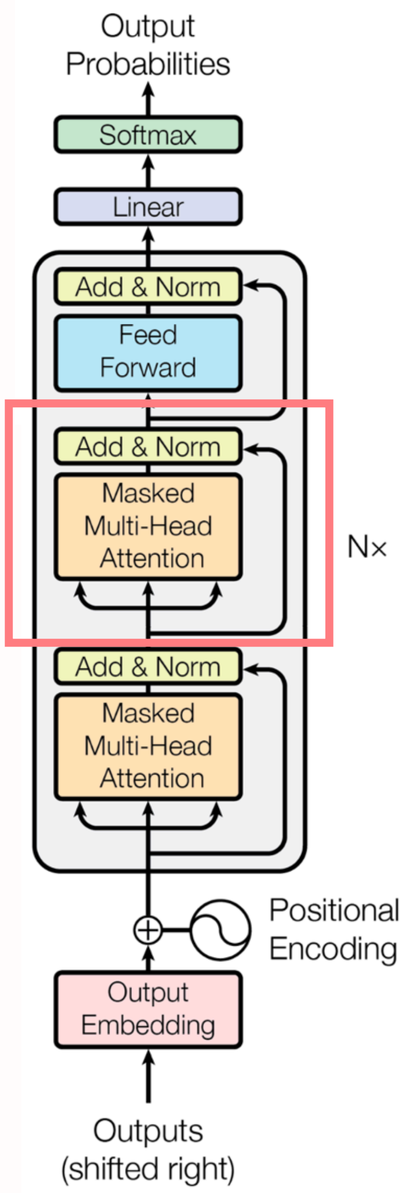
W_p = position embedding matrix

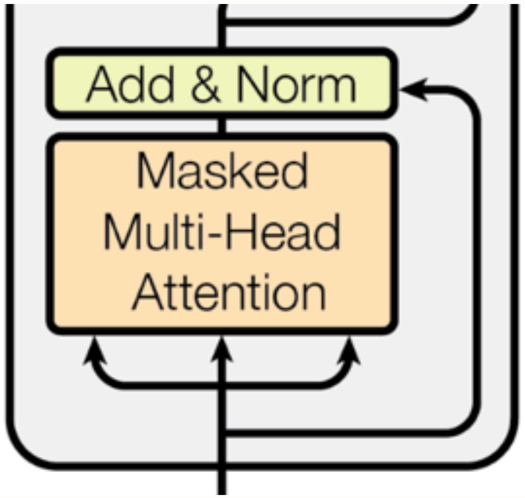
* Byte Pair Encoding

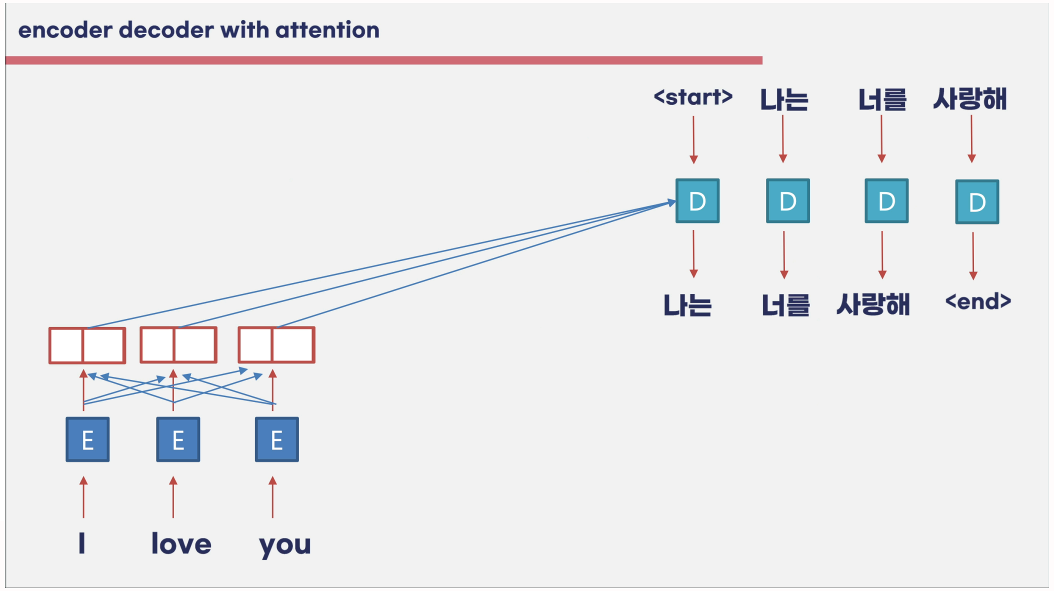
- character 단위로 분리
- vocabulary 로 합체
- 가장 많이 등장하는 유니그램 합체



GPT Unsupervised Pre-Training



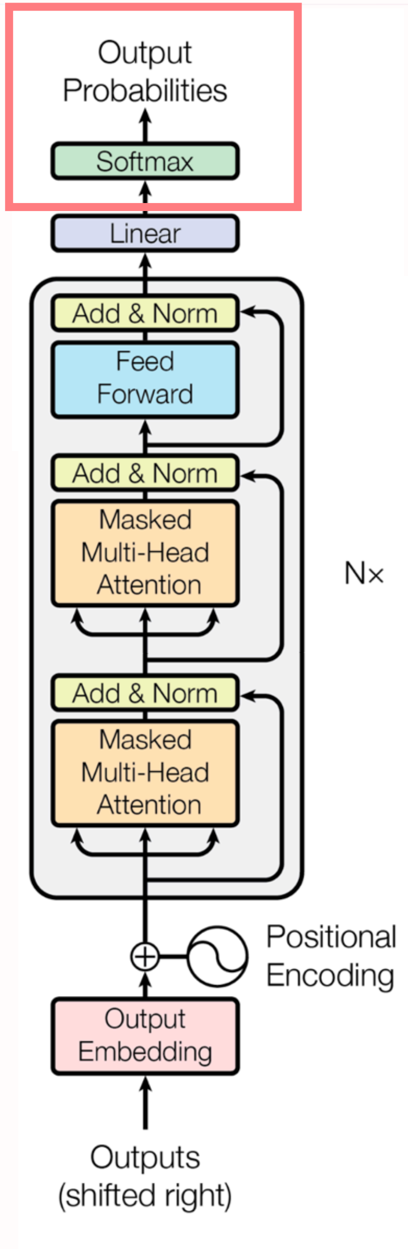


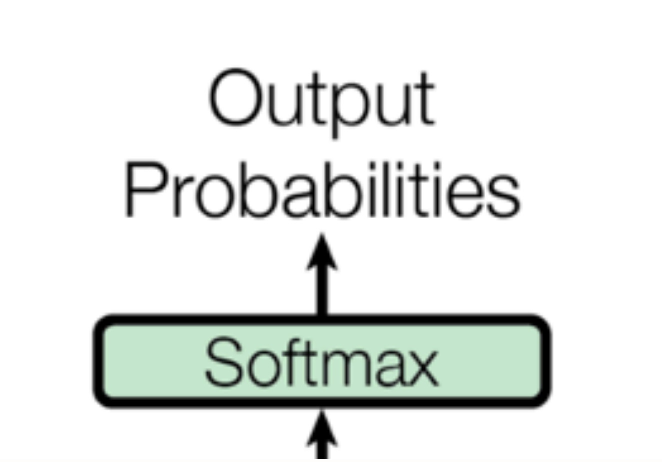


$$h_l = \mathtt{transformer\_block}(h_{l-1})$$

$$\forall i \in [1, n]$$

GPT Unsupervised Pre-Training





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

GPT Supervised Fine-Tuning

labeled Text
(Specific task)

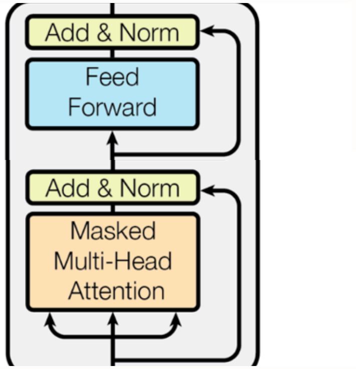
$$C = \{x^1, x^2, \dots, x^m \mid y\}$$

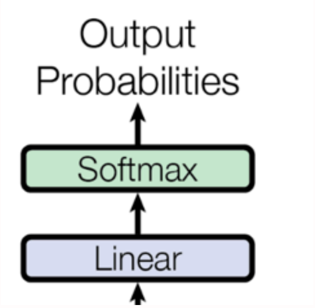
GPT LM

pre-trained model

마지막 transformer block의 h_l

$$h_l^m$$





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

⇒ token들을 linear 결합해서 softmax 적용

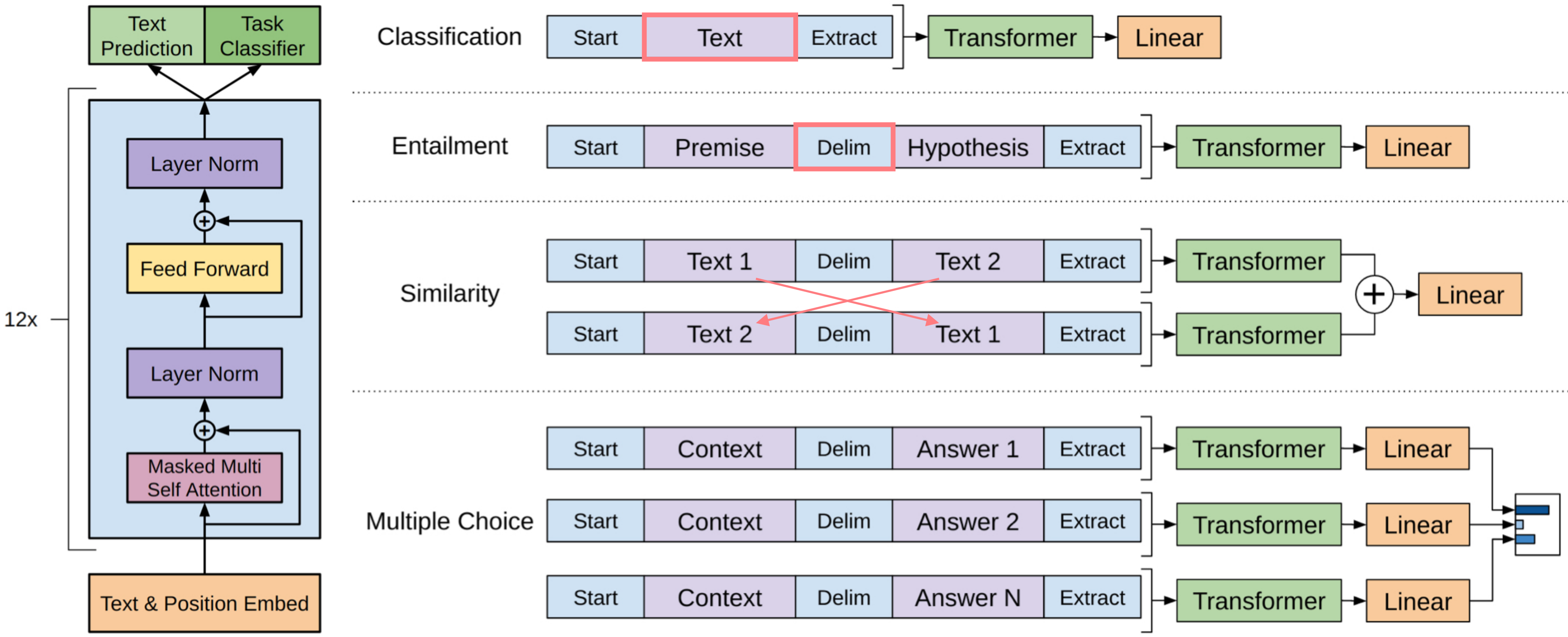
이에 따른 목적함수 $L_2(C)$ 최대화

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

보조 objective 사용? generalization & 학습 속도 향상

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

GPT Task-specific Input



LM을 만드는 것은 동일

하나의 입력 값으로 표현, 입력 구조만 다름

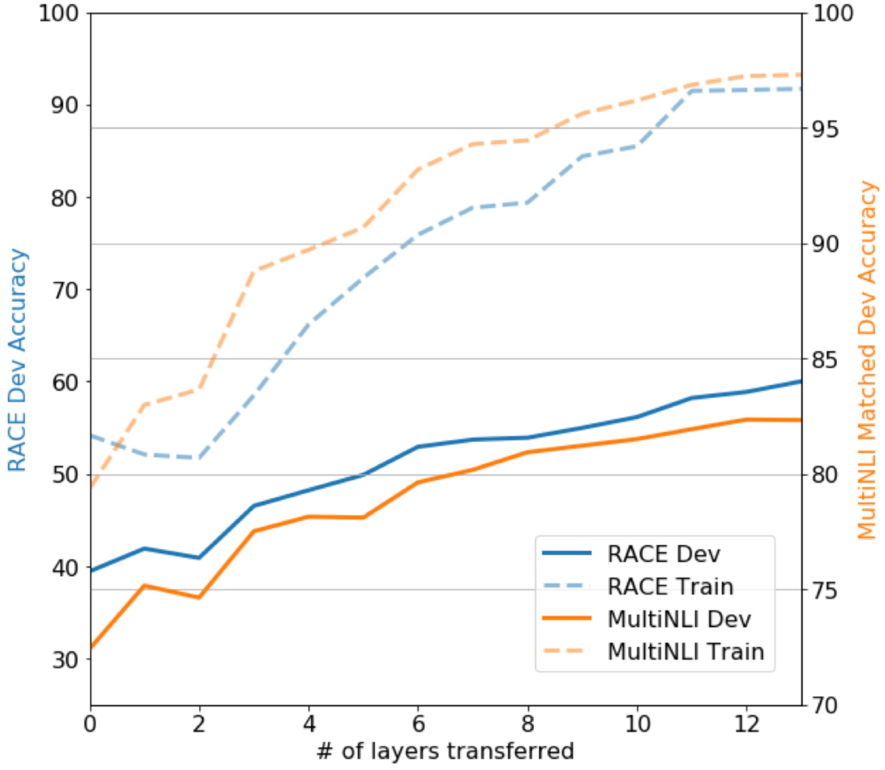
Unsupervised pre-training : BooksCorpus, Word Bechmark 사용

Supervised fine-tuning : 1. 자연어 추론 2. 질의 응답 3. 의미 유사성 4. 분류

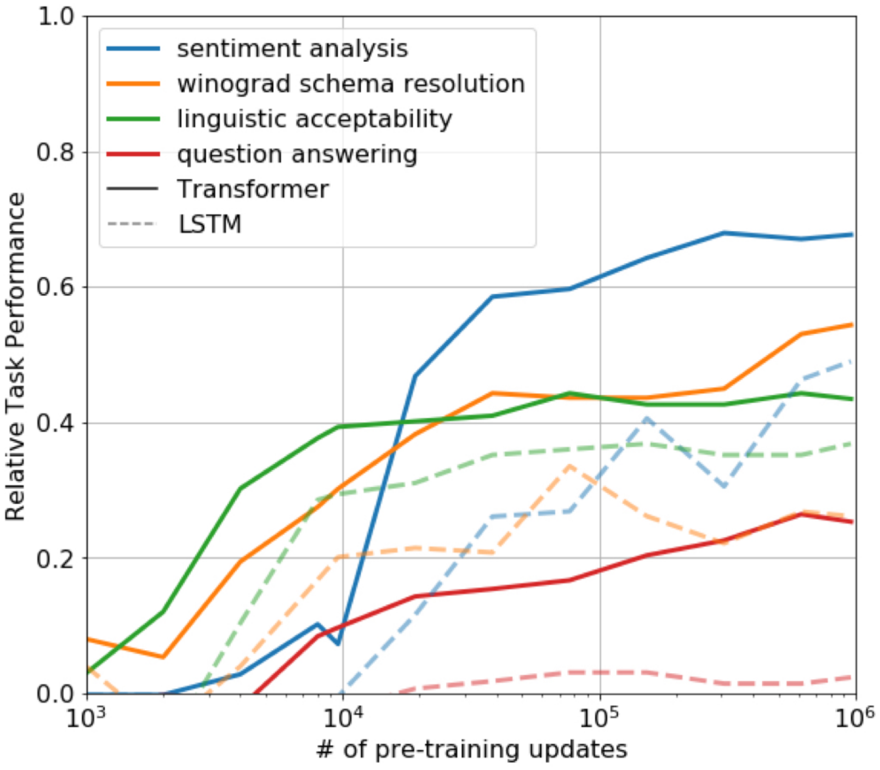
12개의 데이터셋 중 9개에서 state-of-art

GPT Analysis

Transfer layer가 많을수록 성능 향상



fine-tuning으로 파라미터 update 좋음



GPT Analysis

코퍼스 사이즈가 클 때 보조 objective가
성능 개선에 영향을 많이 줌

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

코퍼스 사이즈가 작은 경우, 보조
objective 없이 학습하는 것이 나음

감사합니다

중간고사 화이팅!