GPT-1

임세진

https://youtu.be/5FBZehUCkW4

01. 배경

02. GPT 이해

03. GPT-1 정의 및 구조

04. 성능평가 및 결론

논문 제목 : Improving Language Understanding by Generative Pre-Training

00. 딥러닝 기반의 기계 번역 흐름

- 현재 최신 고성능 기계 번역 모델들은 Transformer 구조를 기반으로 함

- GPT : Transformer의 **디코더(Decoder)** 구조 활용

- BERT : Transformer의 **인코더(Encoder)** 구조 활용

GPT-1 (Generative Pre-Training of a language model)

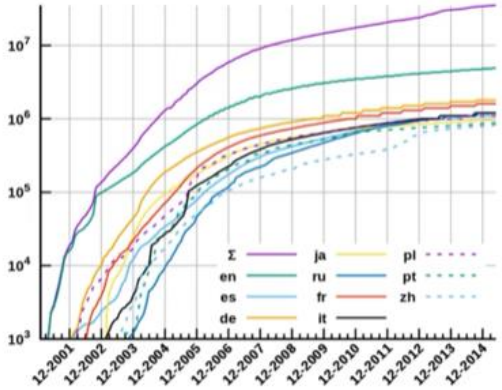
01. 배경

- **데이터의 수는 Unlabeled dataset이 훨씬 많음**

- 지금까지의 모델들은 최적화된 값을 찾기 위해 Labeled dataset을 사용하여 지도 학습을 수행해옴

- 훨씬 많은 Unlabeled dataset을 활용해서 더 좋은 성능을 낼 순 없을까? (훈련에 필요한 시간과 비용 절약)

Unlabeled dataset



Labeled dataset

20.02 기준, English Wikipedia

articles는 600만 건 이상, 35억 개 이상의 단어

01. 배경

- **Unlabeled dataset의 정보를 활용하기 힘든 이유**

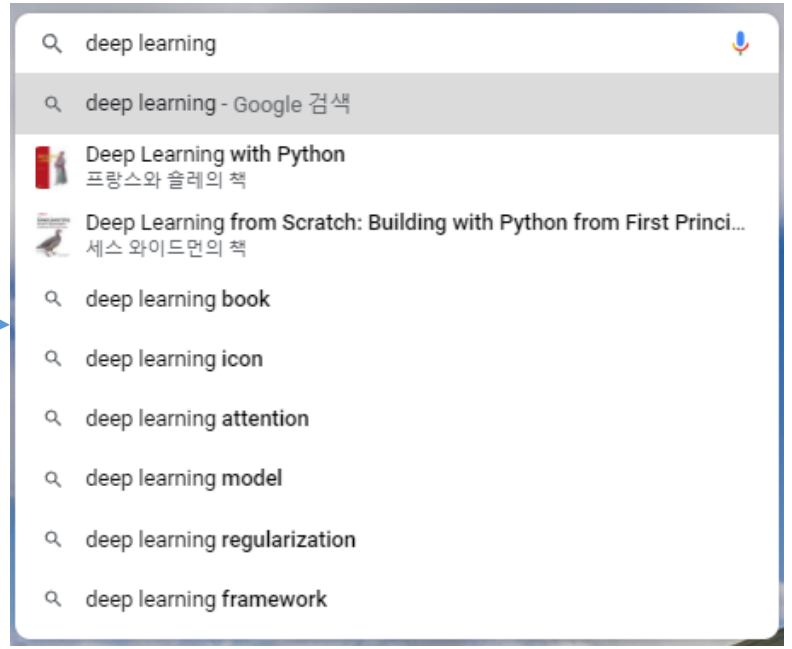
1. 어떤 목적 함수(Optimization objective)가 효과적인지 알 수 없음
2. 모델에서 학습된 표현(representation)을 다양한 NLP task로 전환할 때 가장 효율적인 방법이 정해지지 않음

➔ GPT-1이 이 두가지 단점을 보완

02. GPT 이해

- **Language Model (LM)**

- 단어 시퀀스에 확률을 할당하는 모델 (특정 문장(단어)이 등장할 확률을 계산해주는 모델)
- 이전 단어들을 이용하여 다음 단어를 예측함 (특별히 Labeling이 필요 없음)
- 대량의 학습 데이터로 학습하면 오류율 ↓ , 자연어의 특성을 학습하게 되어 성능 ↑
- 통계를 이용한 방법과 인공신경망을 이용한 방법이 있음
- 기계번역, 음성인식, 철자 교정 등에 응용 가능
- 예) Unigram, n-gram, RNN 계열, Transfomer 계열 등

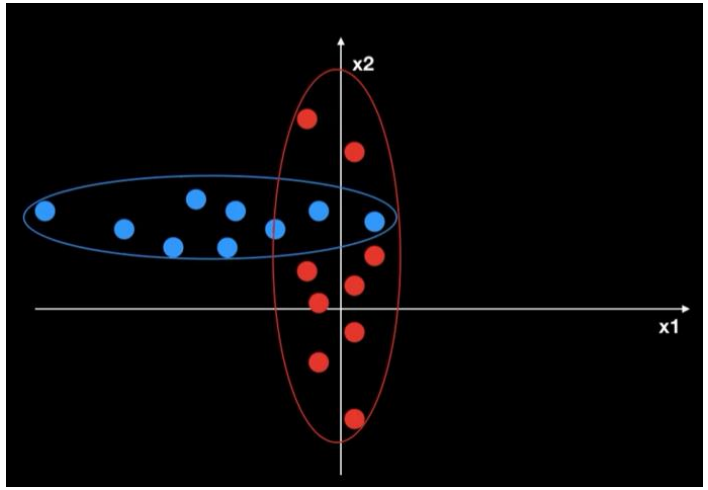


02. GPT 이해

머신러닝 학습 방법 분류

Generative model

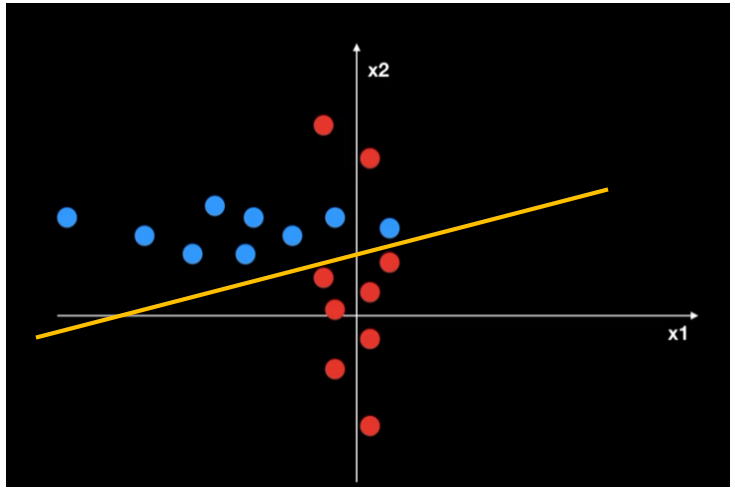
- ✓ 클래스 분류에 분포도 사용
- ✓ 충분한 데이터셋을 가졌을 때 학습효과가 뛰어남
- ✓ overfitting이 상대적으로 적게 발생
- ✓ 상대적으로 연산이 많음
- ✓ Language Model의 학습 방법



VS

Discriminative model

- ✓ 두 클래스에 차이에 초점
- ✓ 적은 데이터로도 좋은 성능을 보임
- ✓ 상대적으로 연산이 적음
- ✓ 한정된 데이터셋에 overfitting 되기 쉬움
- ✓ 주로 많이 사용되는 학습 방법



02. GPT 이해

- **Pre-training (사전 학습)**
- **GPT-1은 단순한 LM이 아니라 아래 유형의 문제에서도 뛰어난 성능을 보여줌**

- Natural Language Inference
- Question Ansewring
- Semantic Similarity
- Classification

✓ GPT의 목적은 대량의 dataset을 이용해서 자연어 처리 능력이 뛰어난 모델을 만드는 것 !

✓ 원래 NLP를 응용하려면 task 관련 layer를 추가적으로 연결해야함

✓ GPT-1은 layer 추가없이 추가 학습(fine-tuning) 가능. 모델이 이미 자연어 처리 능력이 뛰어나므로!

03. GPT-1 정의 및 구조

- **Transformer (2017, Google) : 기계 번역**

- **GPT-1 (2018, OpenAI)**

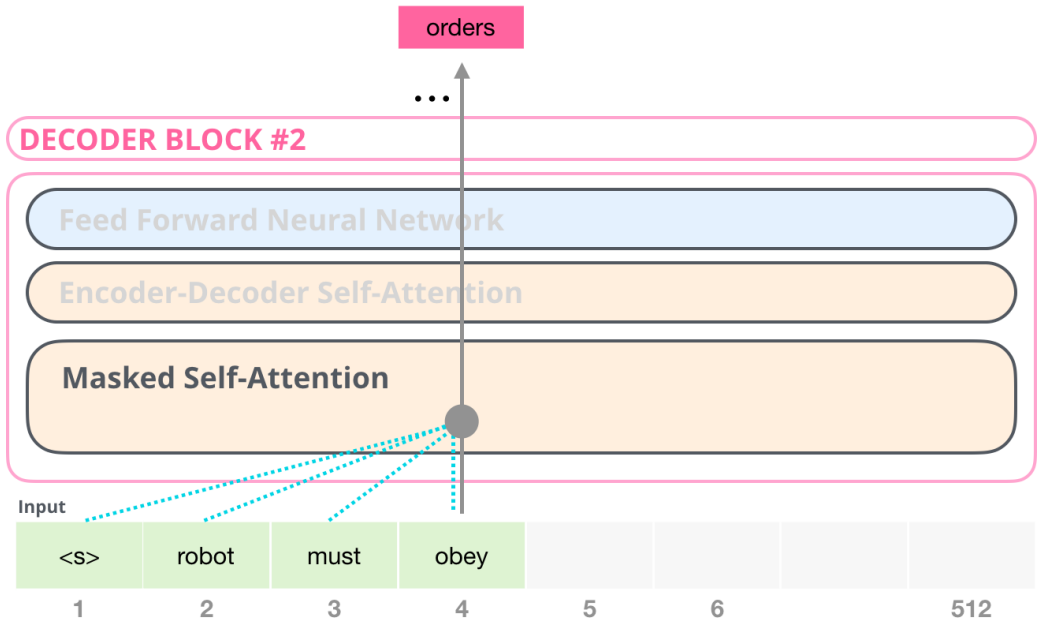
  - Natural Language Inference (두 문장의 관계 유추, entailment / contradiction)
  - Question Answering (질의응답, 정보가 담긴 문장과 질문을 줬을 때 알맞은 응답을 하는지)
  - Semantic Similarity (비슷한 문장 맞히기)
  - Classification (분류)

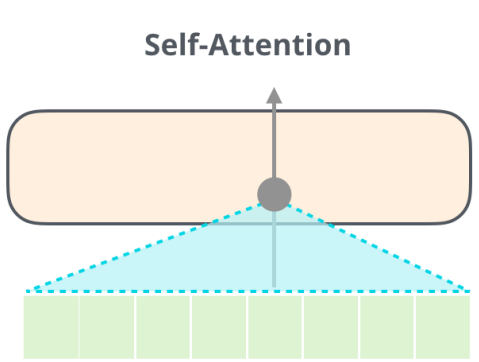
Sentence1	Label
I am happy	Positive
I am sad	Negative

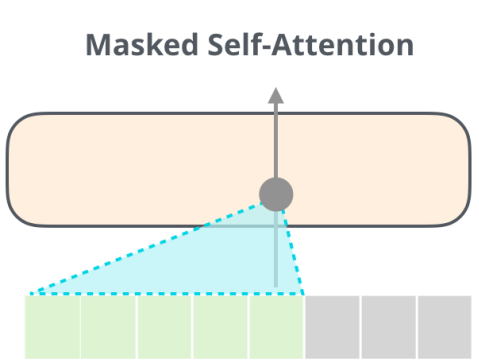
Sentence1	Sentence2	Label
남자는 동아시아 국가에서 유니폼의 수치를 점검한다.	그 남자는 자고 있다.	Contradiction (모순)
여러 명의 남자들이 축구 게임을 하고 있다.	몇 명의 남자들은 운동을 하고 있다	Entailment (참)

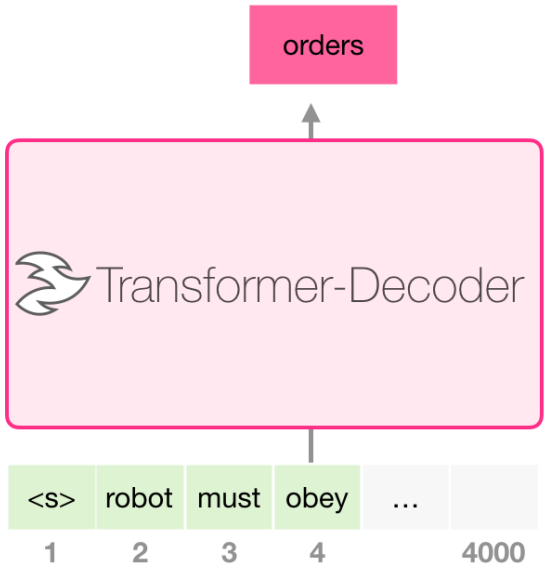
03. GPT-1 정의 및 구조

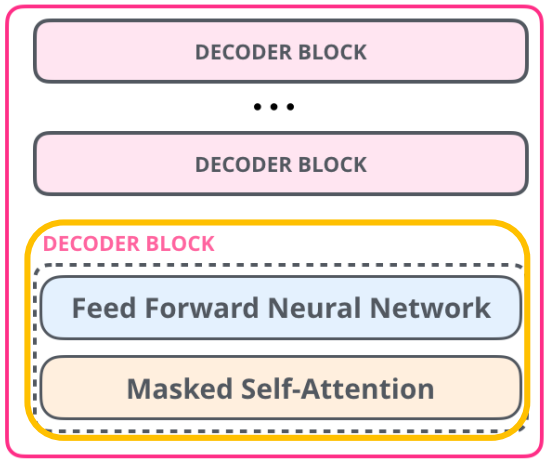
Transformer의 Decoder 구조를 사용



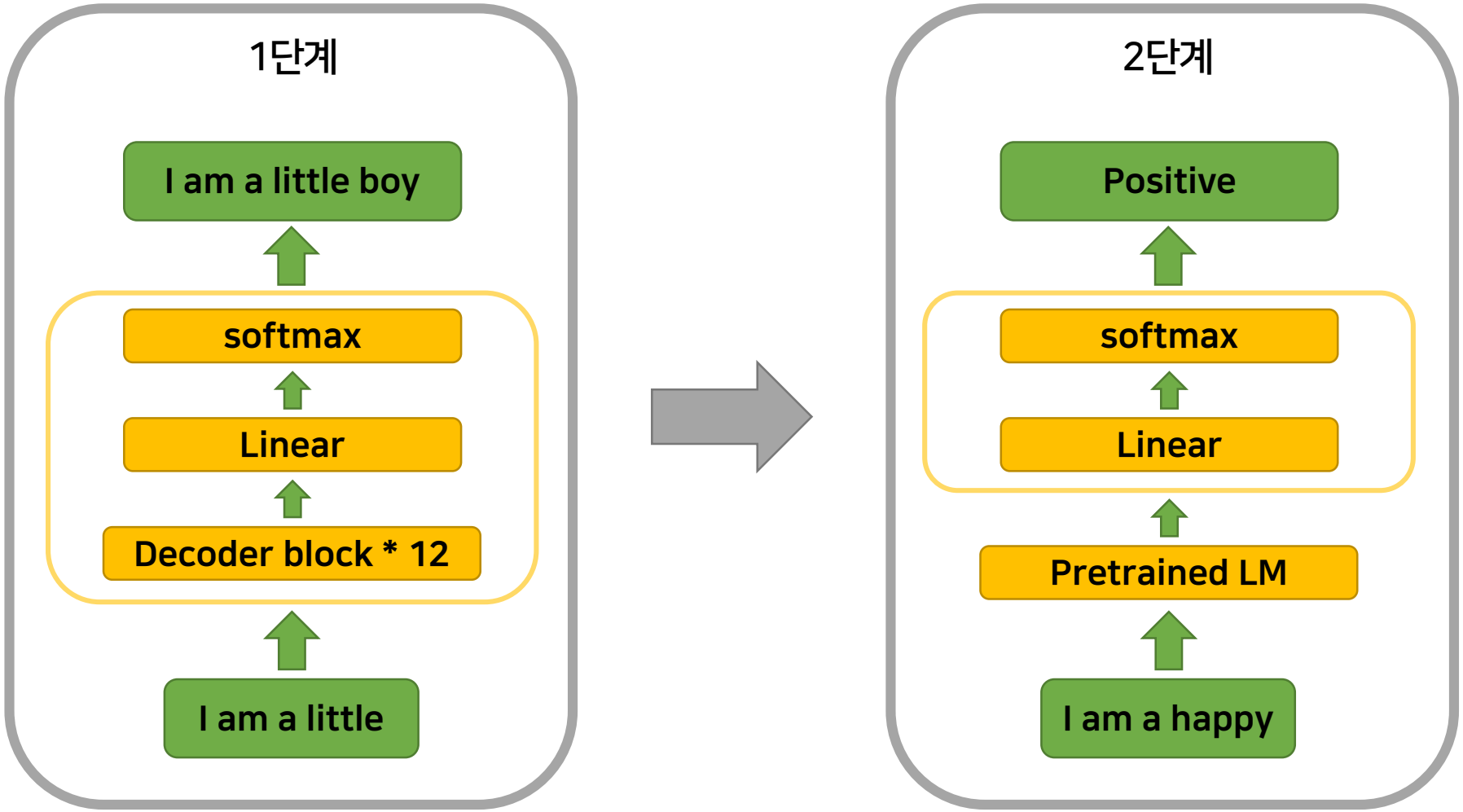








03. GPT-1 정의 및 구조



03. GPT-1 정의 및 구조

- Unlabeled dataset의 정보를 활용하기 힘든 이유

1. 어떤 목적 함수(Optimization objective)가 효과적인지 알 수 없음
2. 모델에서 학습된 표현을 다양한 NLP task로 전환할 때 가장 효율적인 방법이 정해지지 않음

➔ GPT-1이 이 두가지 단점을 보완

- 비지도 학습(unsupervised)으로 사전 학습(pre-training)을 진행하고, 지도 학습(supervised)으로 추가 학습(fine-tuning)을 하여 준지도 학습(semi-supervised)을 구현

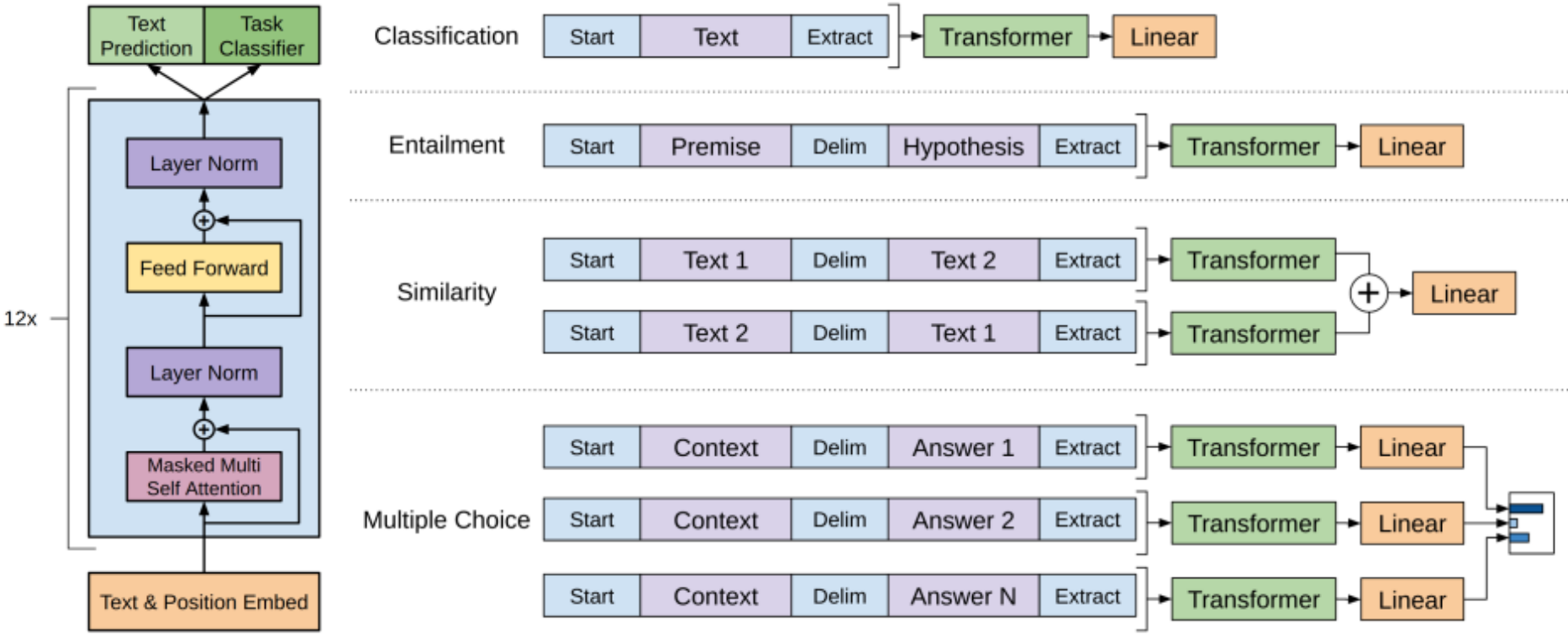
1. 신경망의 초기 매개변수를 학습하기 위해 unlabeled data에 언어 모델링 목적 함수를 사용
2. 앞에서 얻은 parameter를 supervised objective를 사용하여 fine-tuning한 후, 특정 task에 적용

03. GPT-1 정의 및 구조

1. Unsupervised Pre-training with LM objective function (LM 학습) : Pre-Training

2. Supervised Fine Tuning (새로운 Layer 추가없이 모델 자체에서 이어서 학습) : Fine Tuning

- Labeled data를 입력하여 최적화만 하면 됨
- Task에 따라서 입력의 형태가 달라짐

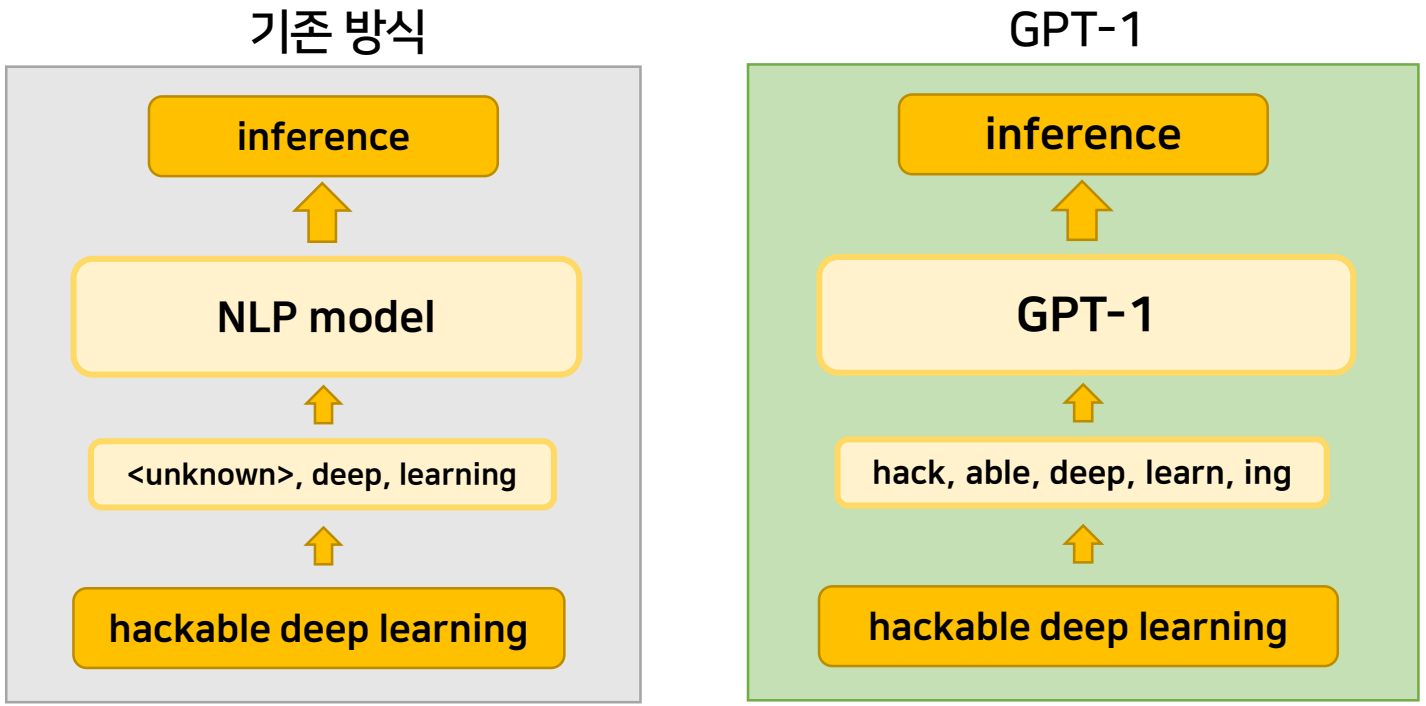


03. GPT-1 정의 및 구조

- **진화된 임베딩 방법 : BPE (Byte Pair Encoding)**

기존 임베딩 방법

- Word embedding : 단어 간 유사도를 찾기 쉬우나, 학습 데이터에 없는 단어는 유사도가 제로 벡터임 ➔ 신조어와 오탈자에 취약
- Character embedding : 제로 벡터가 나올 확률은 매우 낮으나, 단어 간 유사도가 word embedding에 비해 떨어짐



04. 성능평가 및 결론

Natural Language Inference

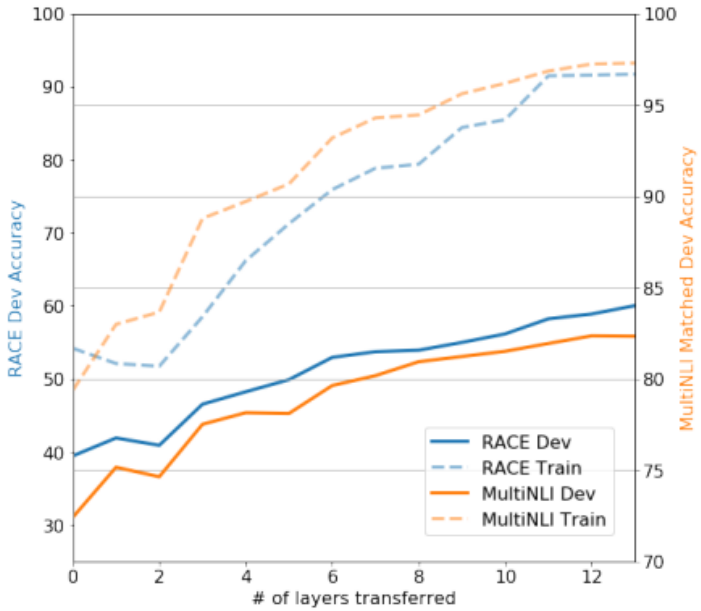
Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Classification & Semantic Similarity

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**

Question & Answering

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**



04. 성능평가 및 결론

- GPT-1

1. Transformer의 decoder 기반 모델
2. 사전 학습은 Unlabeled text를 데이터셋으로 하여 비지도 학습으로 이루어짐
3. 추가 학습은 task specific model을 더하지 않고 그대로 진행한다.
4. Byte Pair Encoding을 사용하여 임베딩을 업그레이드함

Q & A