

# **Week3. NLP**

## **models 3-4. GPT, ALBERT,**

### **RoBERTa**

신승진

## **Contents**

1. Overview
2. GPT

1. GPT2
2. Few-shot Learning
3. ALBERT
4. RoBERTa
5. Reference

## Goal

- ❖ ?  Auto-regression을 설명할 수 있고 GPT 모델 및 BERT 이후에 나온 모델의 특징을 설명할 수 있다.

## 1. Overview

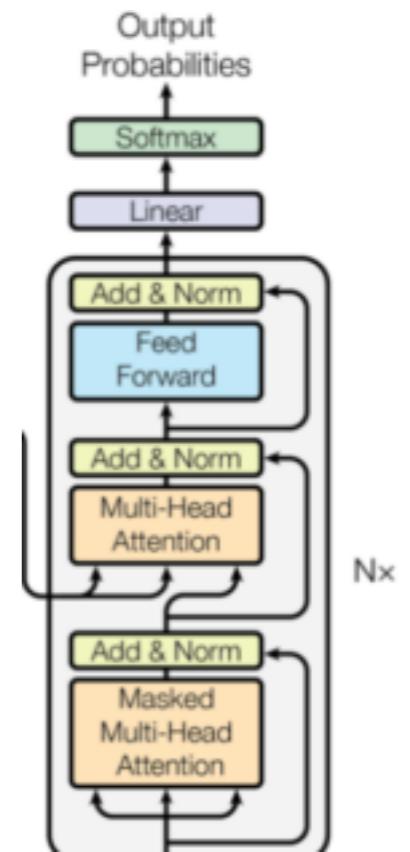
- Auto regression
  - GPT2

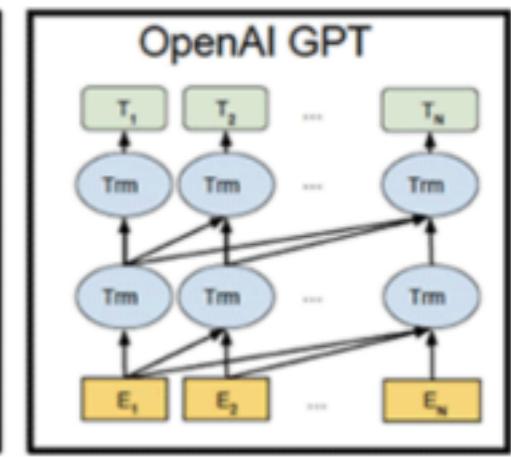
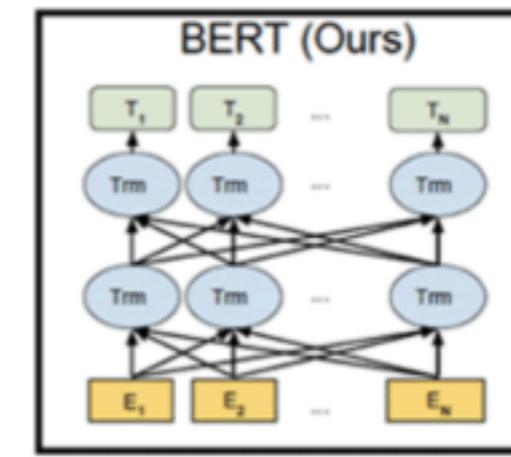
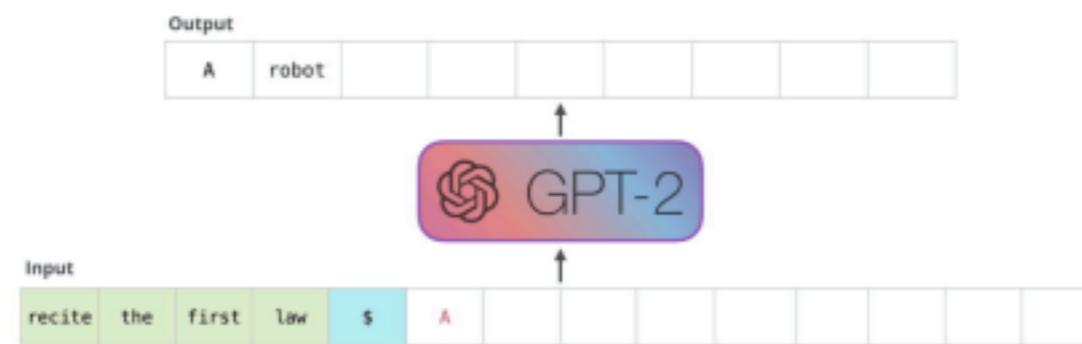
- XLNet
- ...
- Encoder
  - BERT
  - ALBERT
  - RoBERTa
  - Electra
- ...

## 2. GPT

“Generative Pre-trained Transformer”

- Transformer Decoder
- Auto-regressive model
  - Left-to-right transformers





## 2-1. GPT2

**“Language Models Are Unsupervised Multitask Learners”** • Meta Learning

- Multi-task Learning

- GPT 원 모델 아키텍처

- Comprehension

- Summarization

유사

- BPE (Byte Pair

Encoding) • Application

- QA

Supervised Data X (zero-shot setting)

- Machine Translation

\* parameter 업데이트 X

- Reading

$y_h$

## 2-2. Few-shot Learning

“Language Models Are Unsupervised Multitask

# Learners”

Supervised Data X (**zero-shot setting**)

\* parameter 업데이트 X

$y_h$

## 3. ALBERT

### “A Lite BERT for Self-supervised Learning of Language Representations”

- 기존 문제점
  - 모델 규모가 커지면 자원의 한계로 상용화가 어려움
  - 모델의 파라미터 수 증가가 성능과 완전 비례하지 않음
- Parameter Reduction (건강한 파라미터 수 다이어트)
  - 수단
    - Factorized embedding parameterization
    - Input embedding dimension < hidden size

- Cross layer parameter sharing
- [성능 향상을 위해] Next Sentence Prediction 목적 함수 → Sentence Order Prediction • 이점

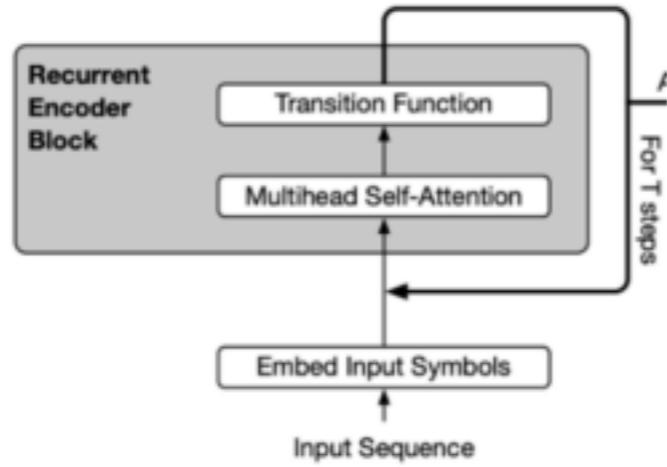
- Memory 사용량 ↓
- Train Time ↓

## 3. ALBERT

“**A Lite BERT** for Self-supervised Learning of Language

**Representations** • Factorized embedding parameterization

- Input embedding dimension < hidden size
- Cross layer parameter sharing  
: shared-attention



- Next Sentence Prediction → Sentence Order Prediction

A: “버트의 등장으로 모든 모델은 트렌스포머 구조를 따르기 시작했다.”

A: “버트의 등장으로 모든 모델은 트렌스포머 구조를 따르기 시작했다.”

B: “오늘 저녁은 탕수육을 먹을 것이다.”

B: “엘모는 NLP 연구에서 가장 새로운 모델이다.”

Label: False Label: False

## 4. RoBERTa

“A Robustly Optimized BERT Pretraining

Approach” • BERT 학습을 최적화

- 수단

- 학습 시간 & 배치 사이즈 & 학습 데이터 10x & 데이터 길이 建
- Next Sentence Prediction task 제거
- Masking Pattern 변화
  - 기존 문제: 문장 내 토큰 중 15%를 대체 (80%는 [MASK]로 대체, 10%는 그대로, 10%는 임의 토큰으로 대체) → 매 epoch마다 문장의 패턴이 그대로 유지됨
- 이점
- NLU에 강점

## 5. Reference

- [2021 SOTA NLP 모델 설명](<https://www.topbots.com/leading-nlp-language-models-2020/>)



- [Illustrated GPT 2](<https://jalammar.github.io/illustrated-gpt2/>) ♦? ♦? ♦? ★★★ ●  
[ALBERT 설명 블로그](<https://y-rok.github.io/nlp/2019/10/23/albert.html>) ♦? ♦? ♦? ★★  
★ ●  
[GPT]([https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf)) ♦? ♦? ♦? □□□
- [GPT 2]([https://d4mucfpksywv.cloudfront.net/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)) ♦? □
- [GPT 3](<https://arxiv.org/pdf/2005.14165.pdf>) ♦? □
- [ALBERT](<https://arxiv.org/pdf/1909.11942.pdf>) ♦? ♦? ♦? □□□
- [RoBERTa](<https://arxiv.org/pdf/1907.11692.pdf>) □? ♦? ♦? □□