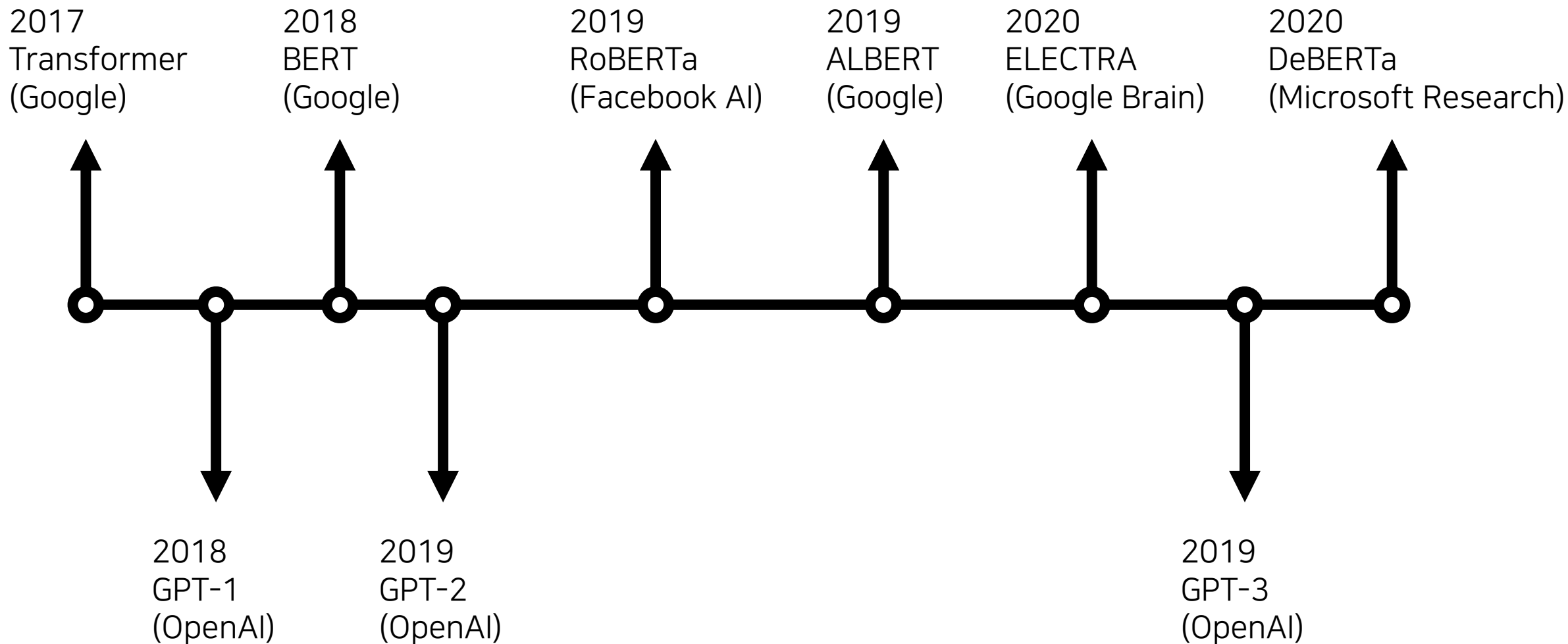


BERT & GPT

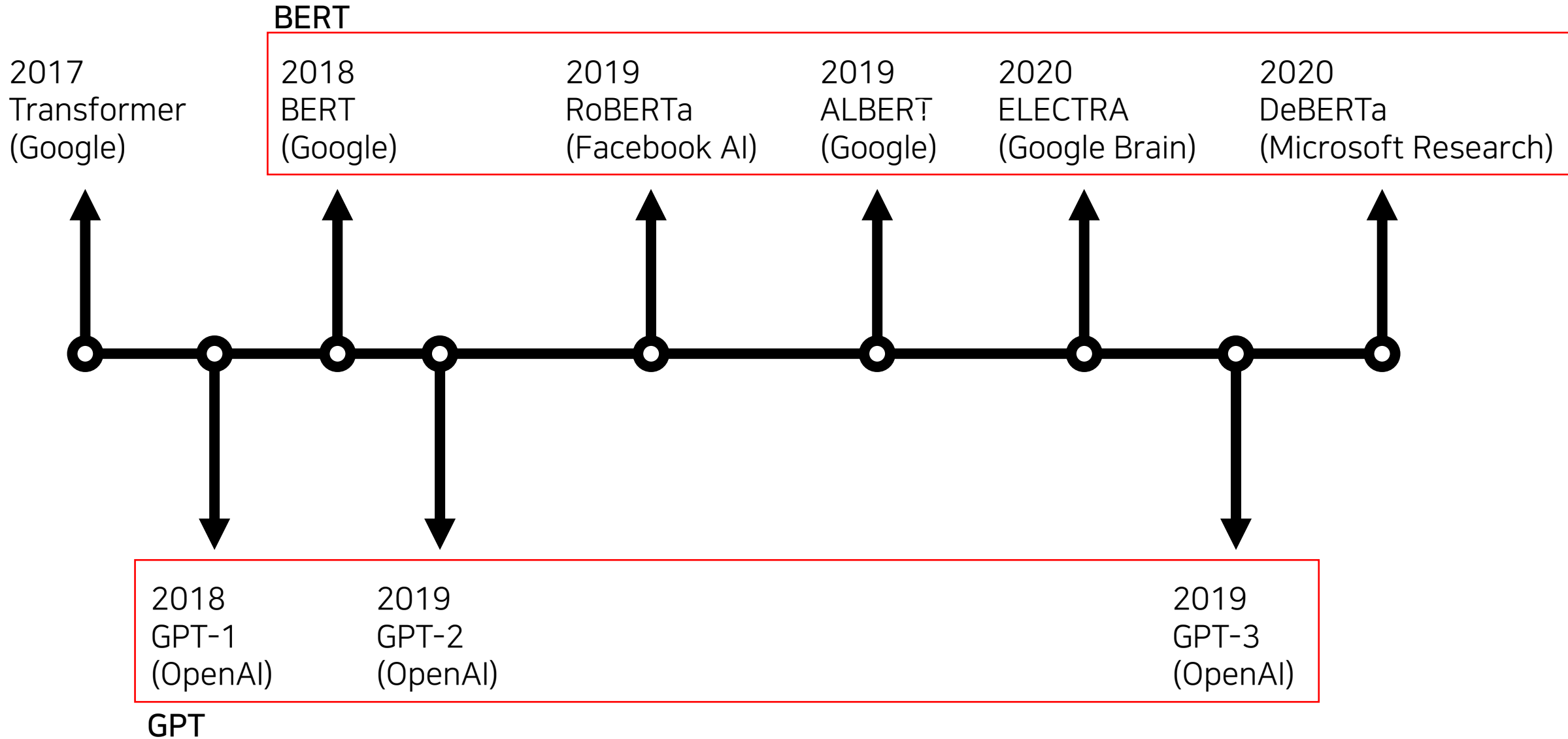
논문 요약 발표

202132033 염지현

Transformer 계보

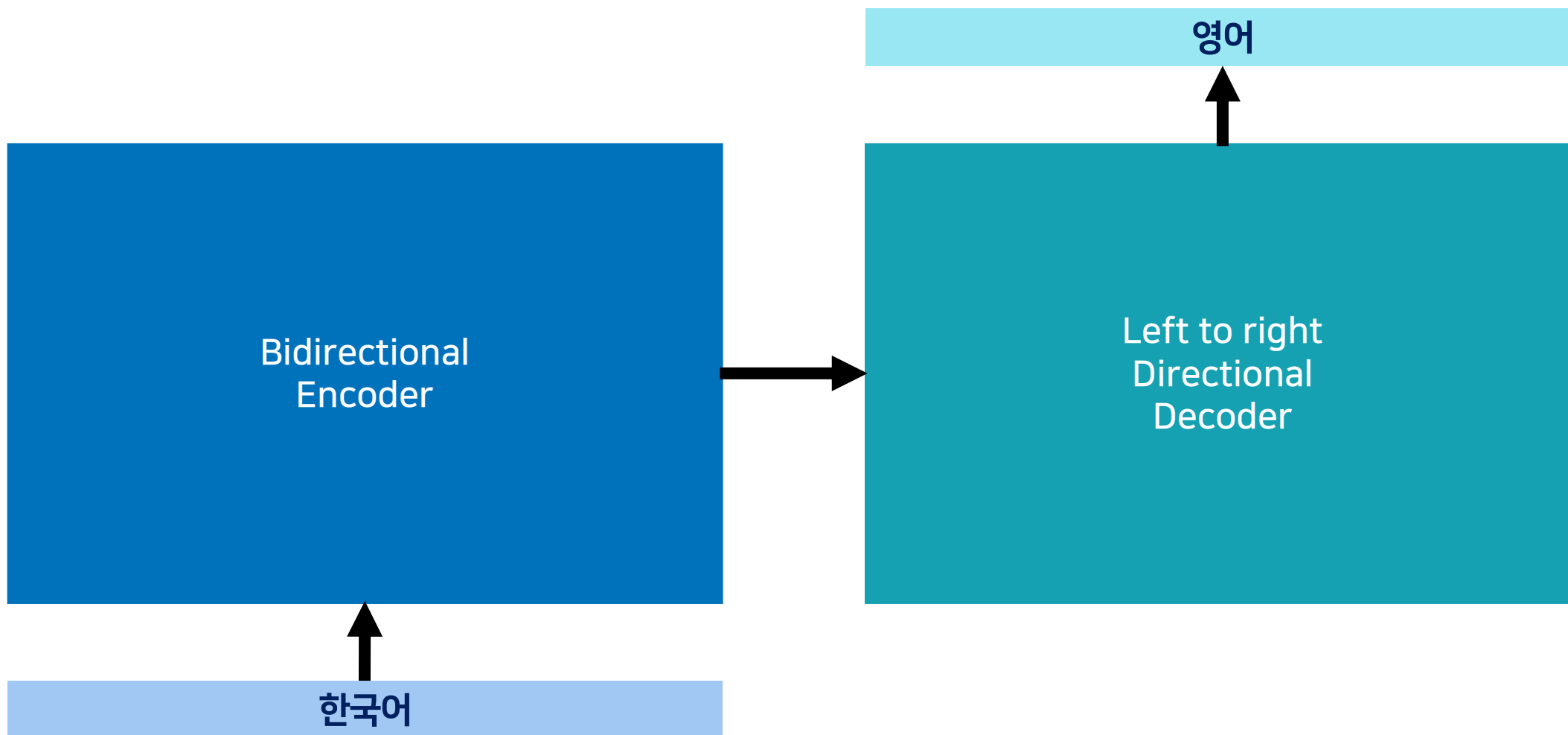


Transformer 계보



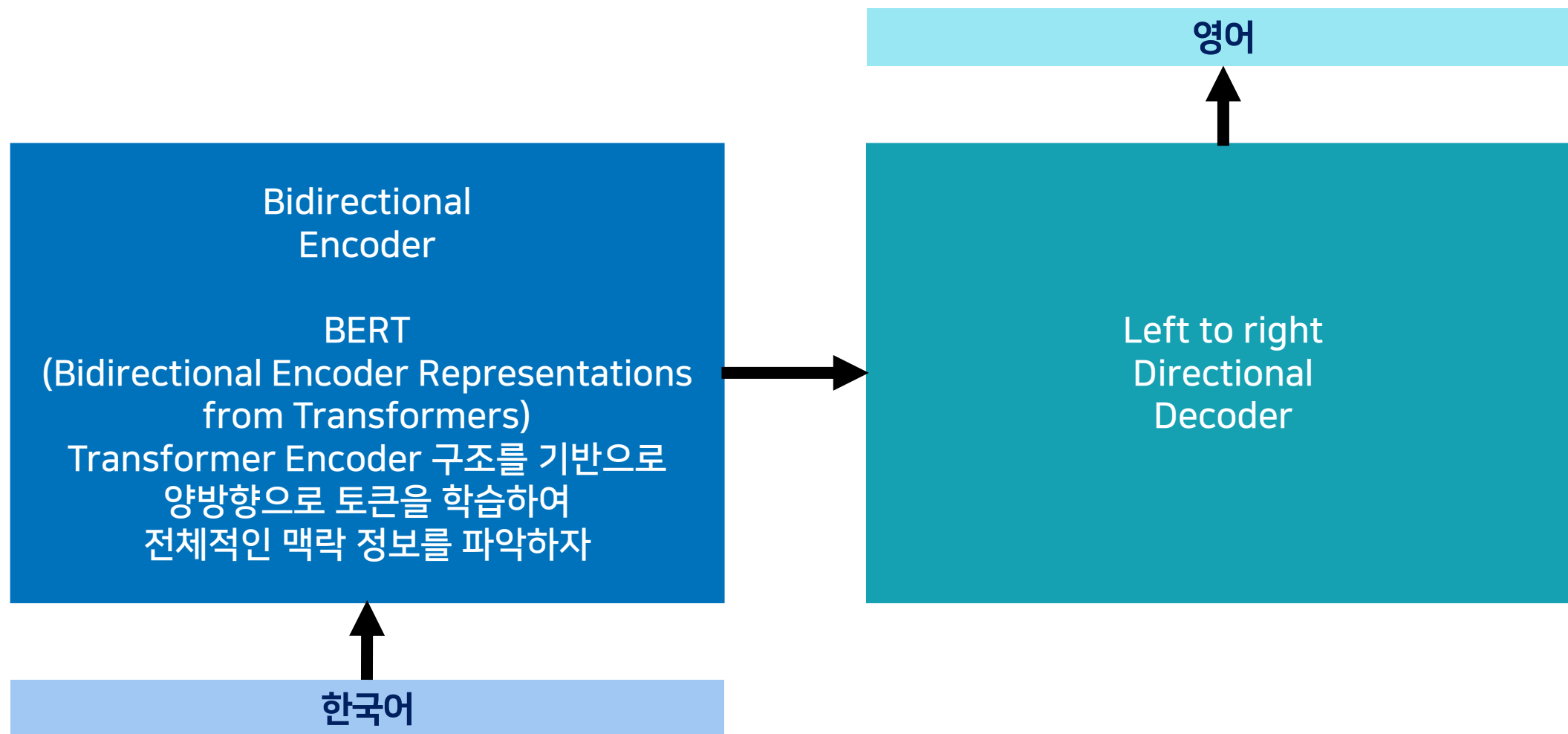
BERT vs GPT

Transformer: 한국어 → 영어 번역 작업일 경우



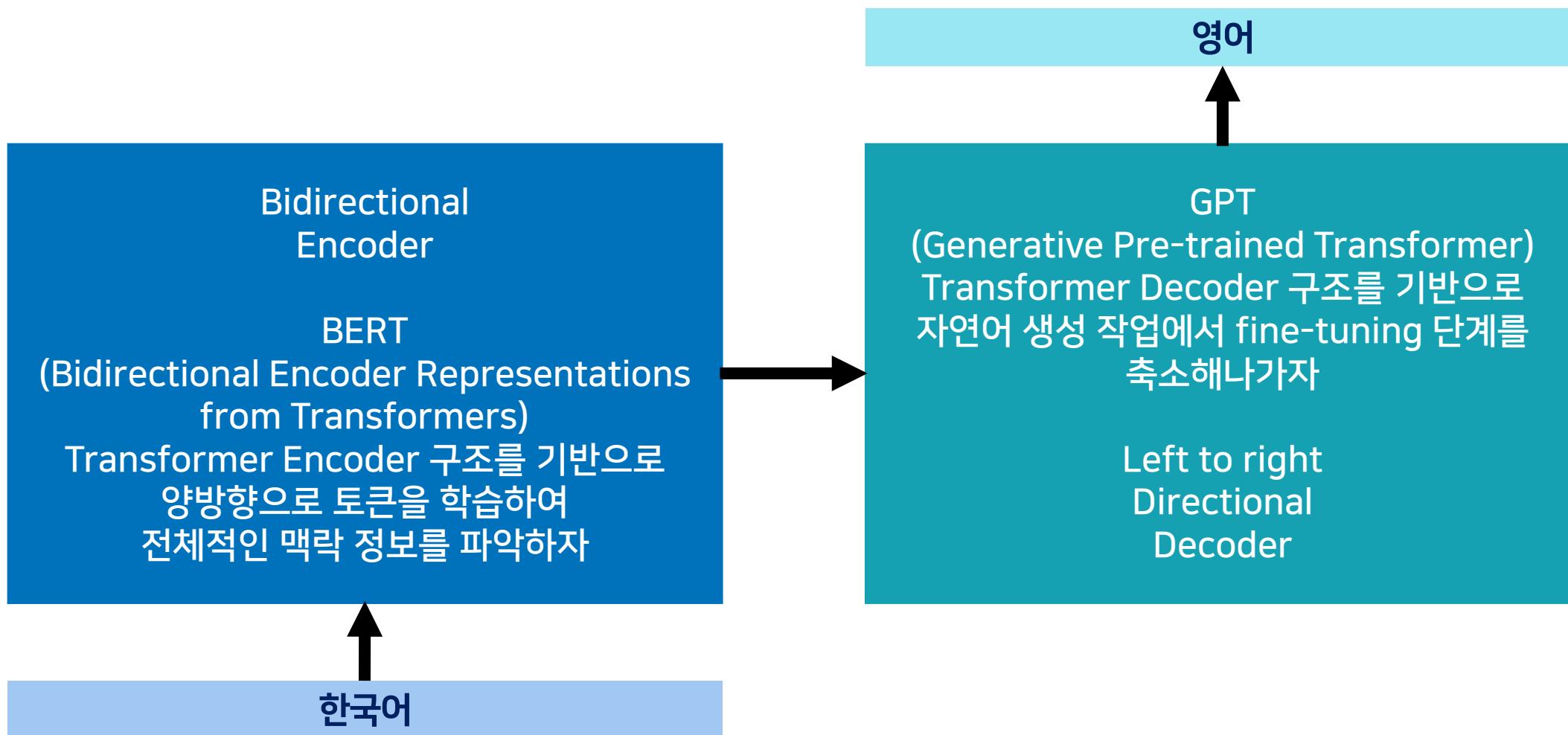
BERT vs GPT

Transformer: 한국어 → 영어 번역 작업일 경우



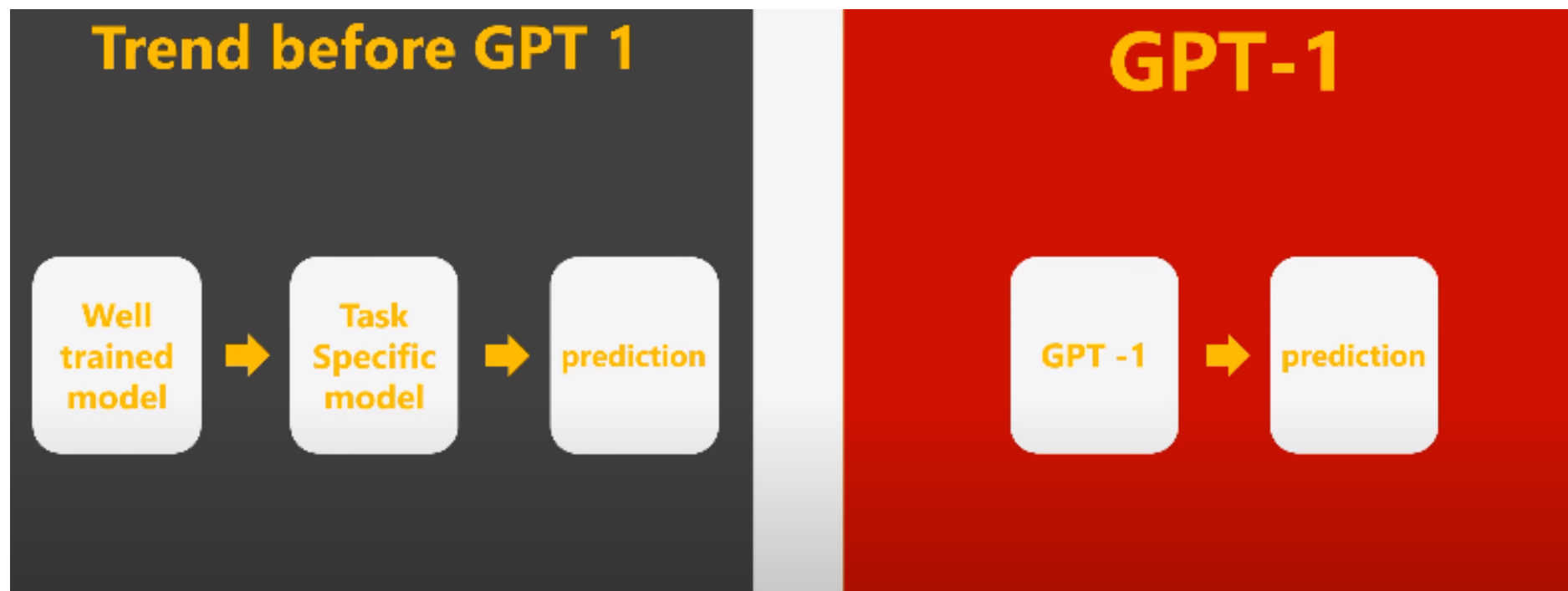
BERT vs GPT

Transformer: 한국어 → 영어 번역 작업일 경우



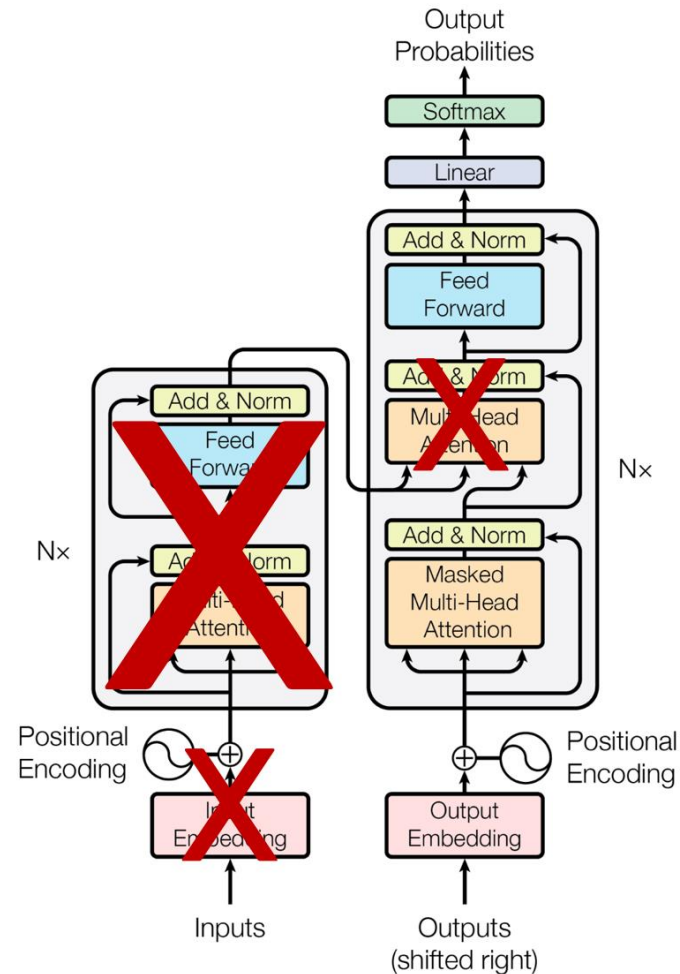
GPT-1: Improving Language Understanding by Generative Pre-Training

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. (OpenAI)



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

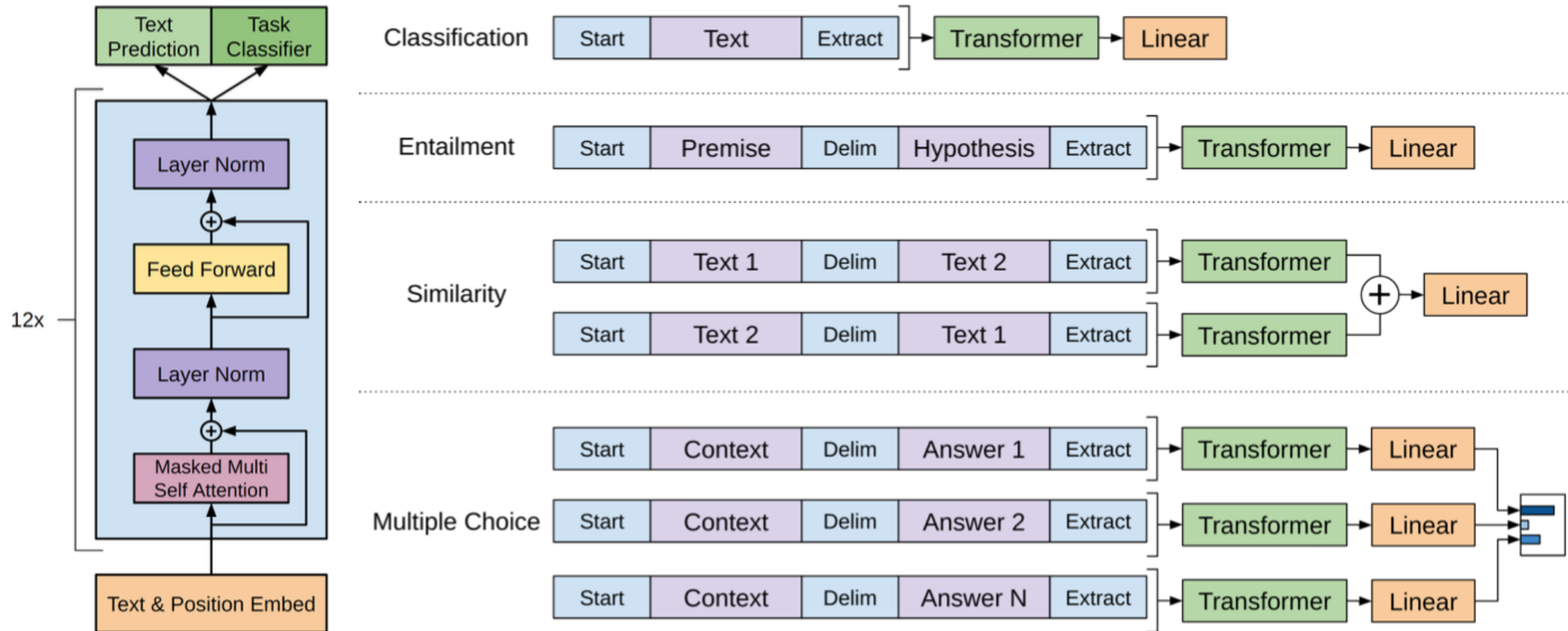
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. (OpenAI)



X 12

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

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. (OpenAI)



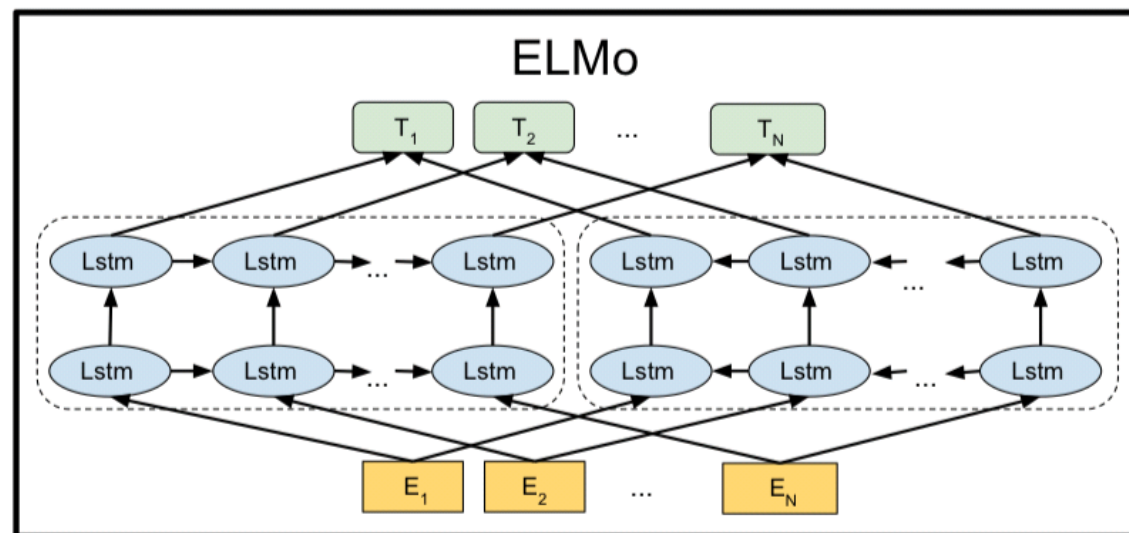
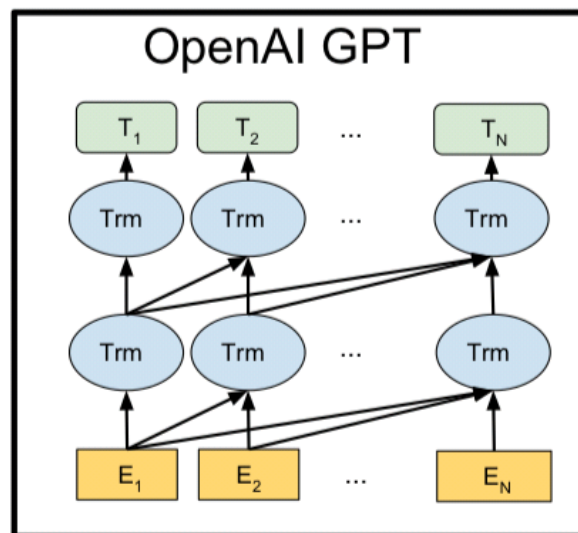
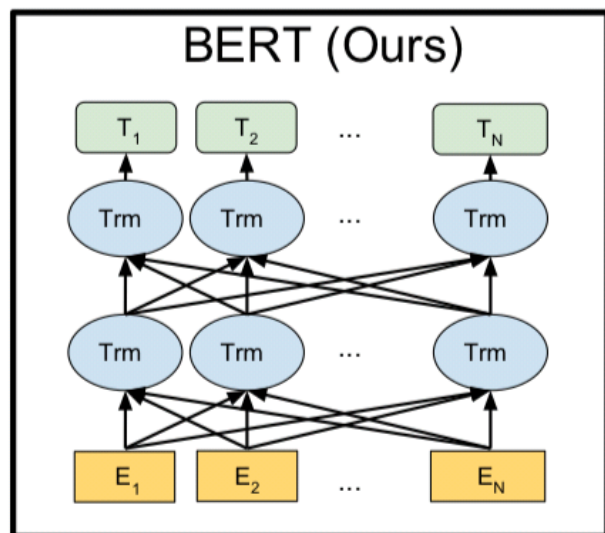
GPT-1: Improving Language Understanding by Generative Pre-Training

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. (OpenAI)

| | |
|----|---|
| 목적 | <ol style="list-style-type: none">1. 기존 연구(pre-training → fine-tuning) 방법에서 fine-tuning 단계는 실제로 미세하지 않음2. Unlabeled text data 활용 |
| 방법 | <p>GPT-1</p> <ol style="list-style-type: none">1. Pre-training: Unsupervised learning → label 이 없는 방대한 양의 데이터를 기반으로 language model 을 학습2. Fine-tuning: Supervised learning |
| 장점 | <ol style="list-style-type: none">1. 따로 fine-tuning을 위한 모델 불필요2. 방대한 양의 데이터로 학습하였으므로 학습 효과 상승3. Unsupervised datasets 사용 가능 |
| 단점 | <ol style="list-style-type: none">1. fine-tuning 단계를 생략하는 대신 큰 모델로 학습을 진행하여 시간과 비용이 많이 듦2. Left to right의 단방향성 학습을 하기 때문에 전체 문맥 정보를 파악하기 한계 존재 |
| 코드 | https://github.com/openai/finetune-transformer-lm |

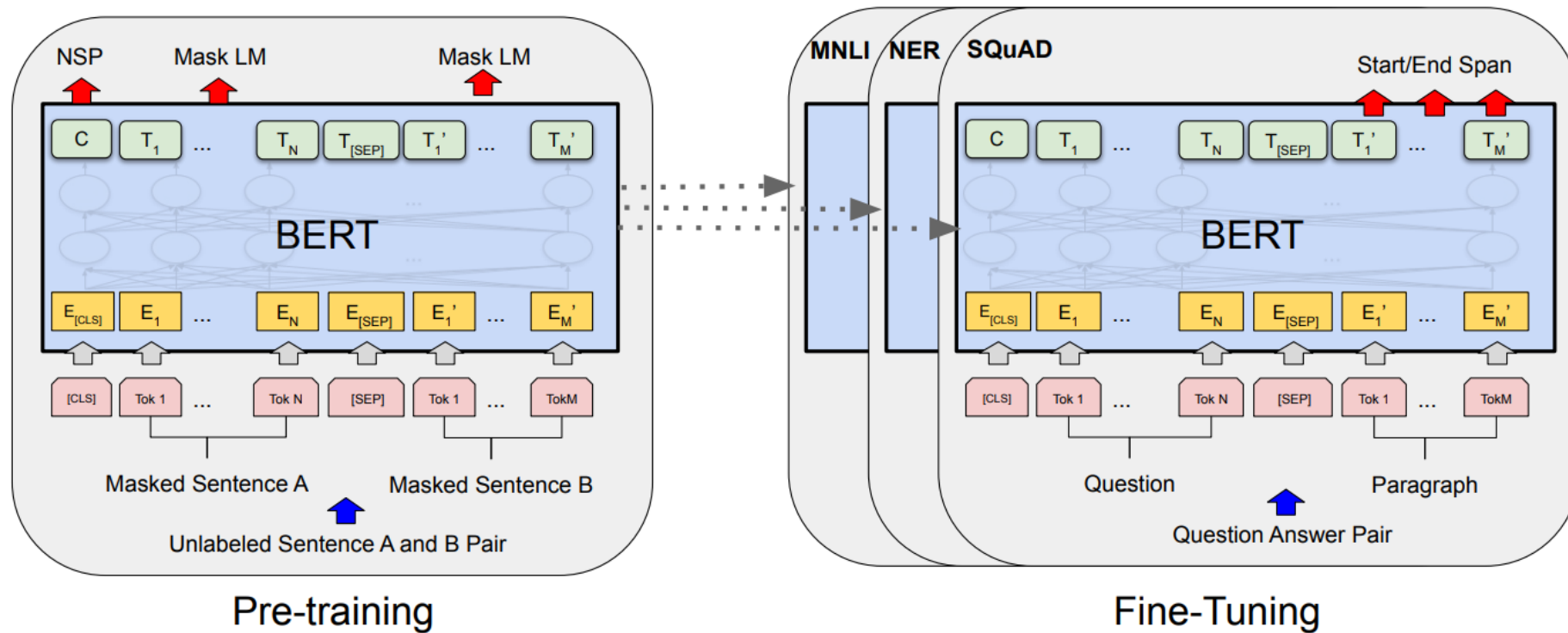
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (Google)



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

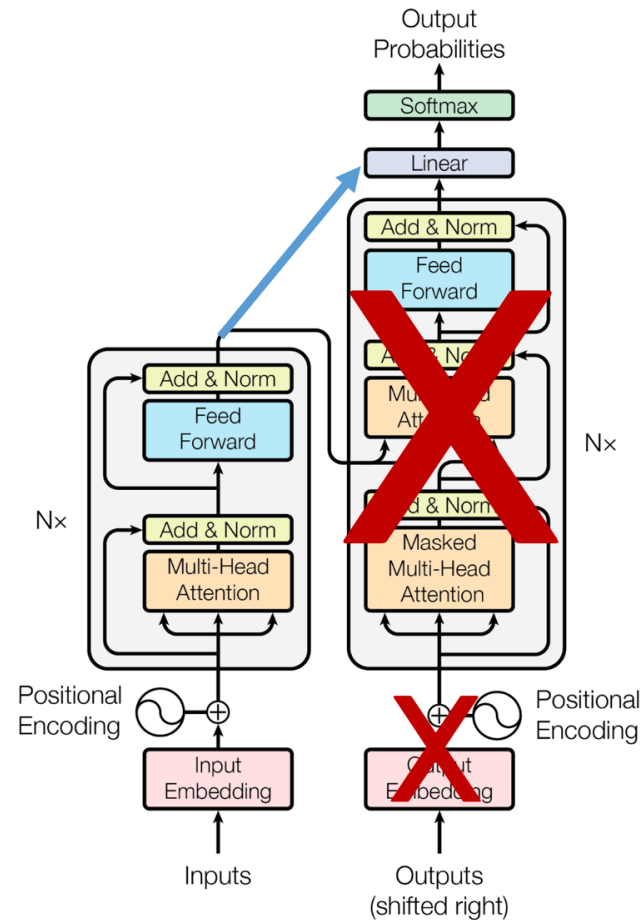
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (Google)



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

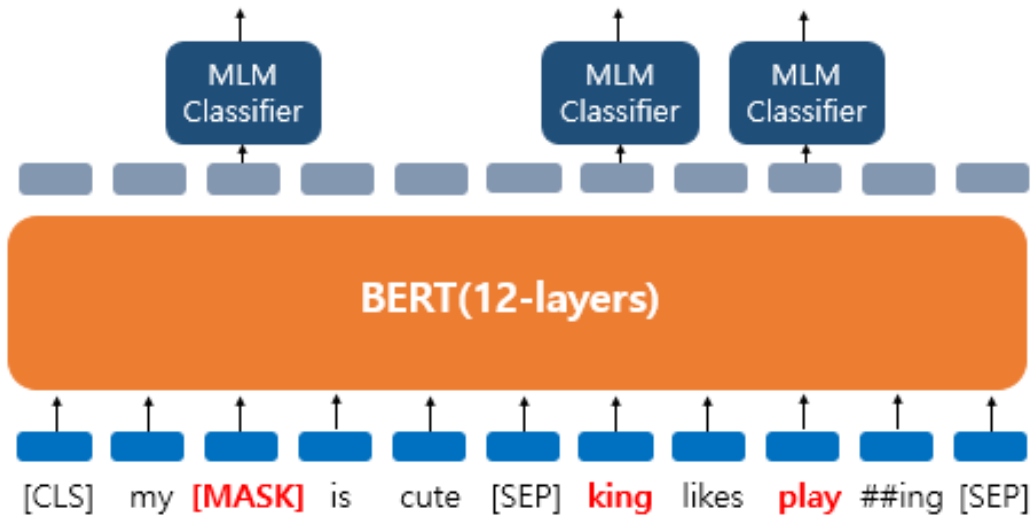
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (Google)

24 X

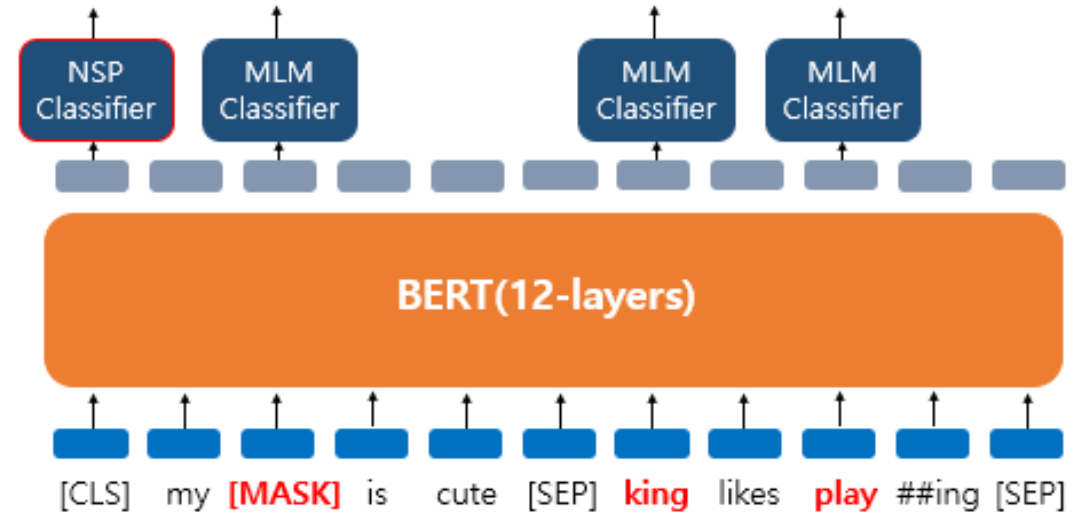


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (Google)



1. MLM(Masked Language Model)



2. NSP(Next Sentence Prediction)

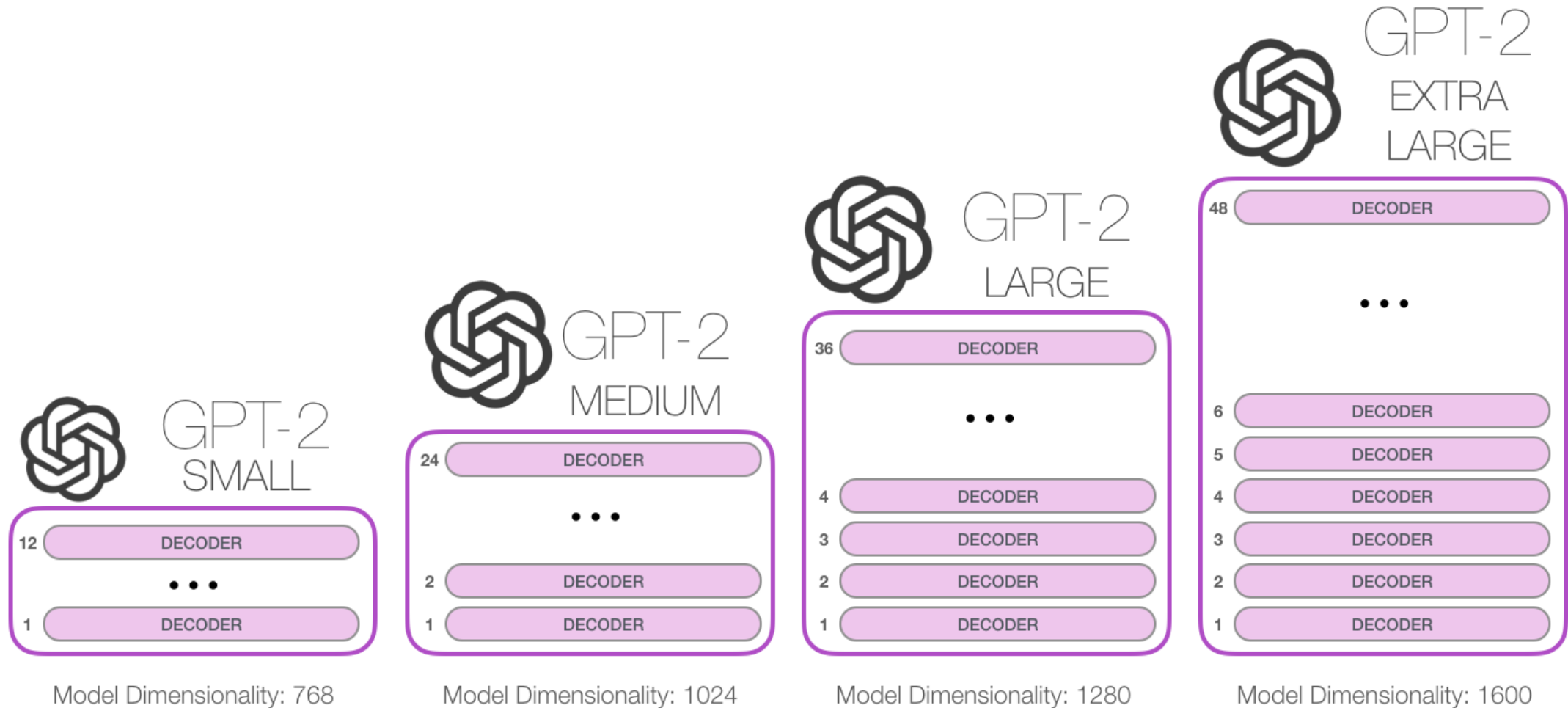
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (Google)

| | |
|----|---|
| 목적 | 기존 단방향 fine-tuning 모델(ELMo, GPT-1)이 이전 토큰만 확인할 수 있는 한계점 개선 |
| 방법 | BERT 1. Pre-training: Unsupervised learning(Contextual embedding) → MLM(Masked Language Model): 문장 중 일부 단어를 [MASK]하여 [MASK] 된 단어를 예측하도록 훈련하기 때문에 각 단어 사이 문맥 정보 조사 가능 → NSP(Next sentence prediction): 두 문장이 이어지는지/이어지지 않는지 예측하는 훈련 2. Fine-tuning: Supervised learning - 각 task별로 데이터세트와 모델 필요 |
| 장점 | 1. GPT-1과 다르게 양방향으로 토큰 사이 연관성을 학습하여 문맥 이해 능력이 비교적 뛰어남 2. GPT-1에 비해 적은 모델로 학습 |
| 단점 | 1. 작업마다 Fine-tuning을 진행해야 하므로 그에 필요한 데이터와 모델 수집 및 학습 필요 |
| 코드 | https://github.com/google-research/bert |

GPT2: Language Models Are Unsupervised Multitask Learners

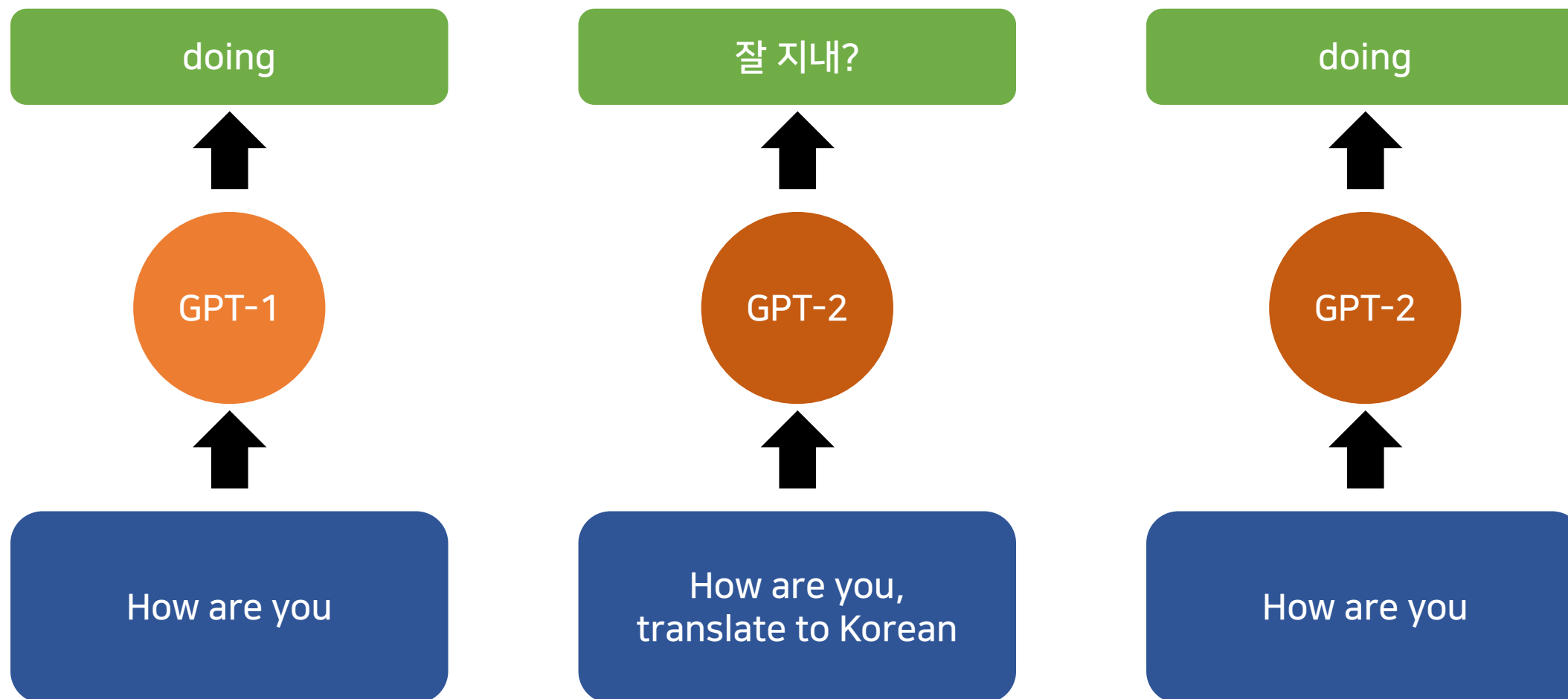
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. (OpenAI)



GPT2: Language Models Are Unsupervised Multitask Learners

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. (OpenAI)

GPT-1: $P(\text{output}|\text{input}) \rightarrow$ GPT-2: $P(\text{output}|\text{input}, \text{task})$



GPT2: Language Models Are Unsupervised Multitask Learners

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. (OpenAI)

| | |
|----|---|
| 목적 | GPT-1(작업에 맞는 데이터 필요), BERT 여전히 fine-tuning 단계 필요 |
| 방법 | GPT-2 1. Zero shot learning 2. 구조는 GPT-1과 동일하지만 학습 데이터와 모델을 더 크게 확장 3. $P(\text{output} \text{input})$ 과 다르게 $p(\text{output} \text{input}, \text{task})$ 와 같이 task를 입력으로 넣음 4. 저자들이 만든 WebText dataset 사용(Reddit에 게재된 글 중 최소 3개의 평가를 받은 글만 수집) |
| 장점 | 1. Zero-shot learning을 기반으로 하는 대규모 NLP 모델 2. 독해 등에서는 supervised baseline 모델과 견줄 만한 성능을 보임 |
| 단점 | 1. 요약 task에서는 성능이 나오지 못함 → 실제 사용 불가 2. QA, Translation 작업에서도 그렇게 좋지 못한 성능을 보이지만 zero-shot learning이기 때 문에 꽤나 인상적인 성능 3. 단방향성 표현의 비효율성 극복 가능성 불명확 |
| 코드 | https://github.com/openai/gpt-2 |

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.(University of Washington, Seattle, WA + Facebook AI)

| | |
|----|--|
| 목적 | BERT 가 상당히 과소 훈련됨(전략적인 훈련 세팅 필요) |
| 방법 | <ol style="list-style-type: none">1. 더 많은 데이터에 대해 더 큰 batch 로 모델을 오래 훈련2. NSP(Next sentence prediction) 사전 훈련 제거<ul style="list-style-type: none">- NSP loss 제거 결과 성능 향상됨3. BERT에서 사용한 시퀀스보다 더 긴 시퀀스로 훈련4. BERT 사전 훈련 MLM에서 masking 단어 패턴을 동적으로 변경<ul style="list-style-type: none">- 기존 BERT처럼 모델 입력 전에 마스킹 하면 동일한 위치의 동일한 토큰만 마스킹 되므로 학습 효율성이 떨어지므로 동적으로 마스킹하여 학습의 효율성 높임 |
| 결과 | 기존 BERT, BERT로부터 아이디어를 얻은 XLNet 등보다 더 좋은 성능에 도달 → 훈련 전략의 중요성 언급 |
| 코드 | https://github.com/facebookresearch/fairseq/blob/main/examples/roberta/README.md |

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations.(Google Research, Toyota Technological Institute at Chicago)

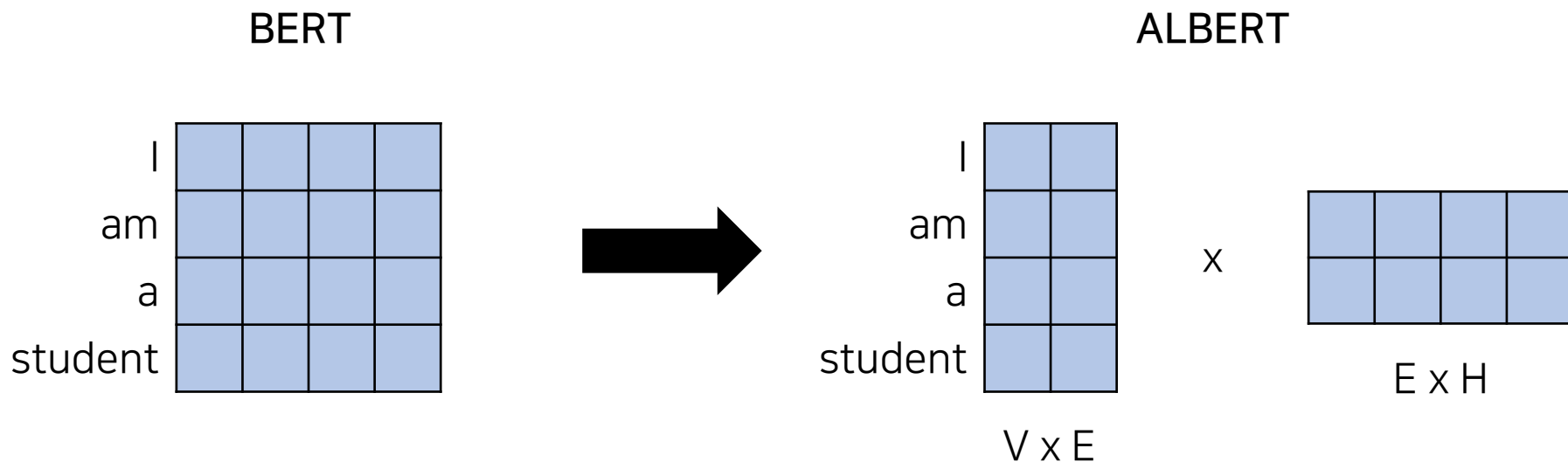
BERT-Base(12 layer) → 110,000,000개 파라미터 사용
BERT-Large(24 layer) → 340,000,000개 파라미터 사용

ALBERT: 이렇게 많은 파라미터가 필요한가?

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations. (Google Research, Toyota Technological Institute at Chicago)

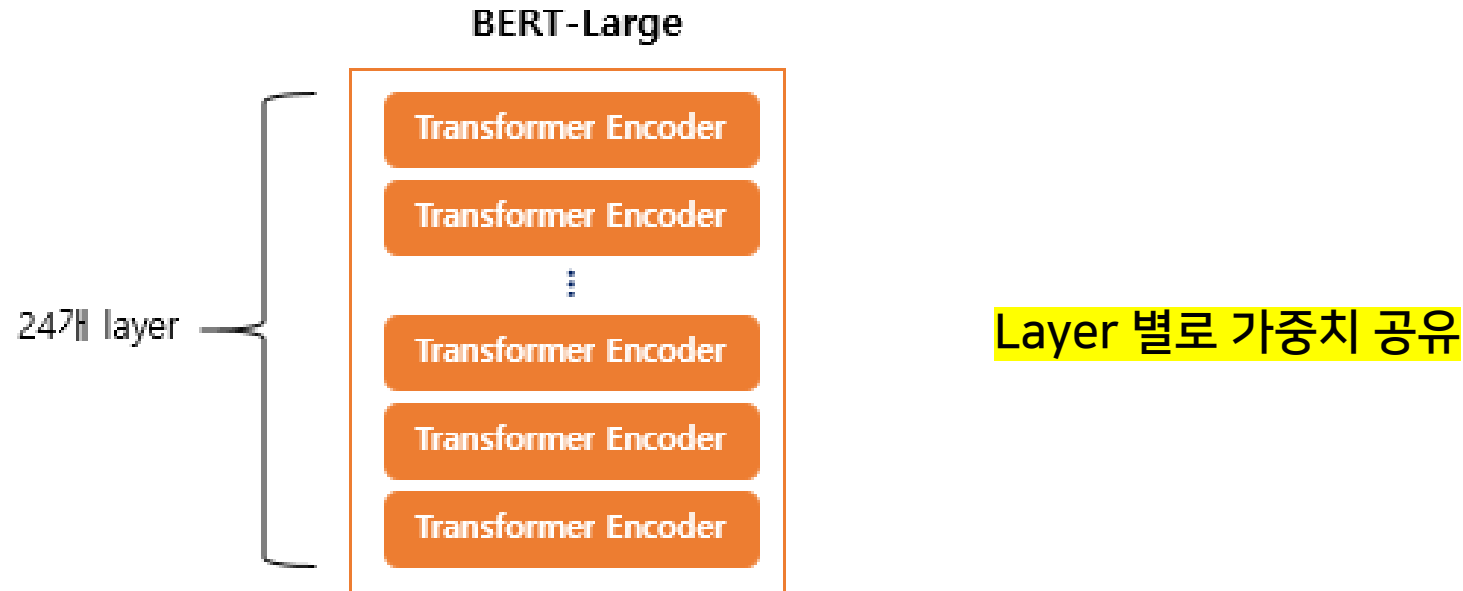
1. Factorized embedding parameterization



ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations. (Google Research, Toyota Technological Institute at Chicago)


2. Cross-layer parameter sharing



ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations.(Google Research, Toyota Technological Institute at Chicago)

3. SOP(Sentence-Order Prediction)



두 문장이 이어지는
문장인지 예측하는
작업

NSP를 대신하는
사전 훈련 작업

두 개의 text
segment 순서 예측

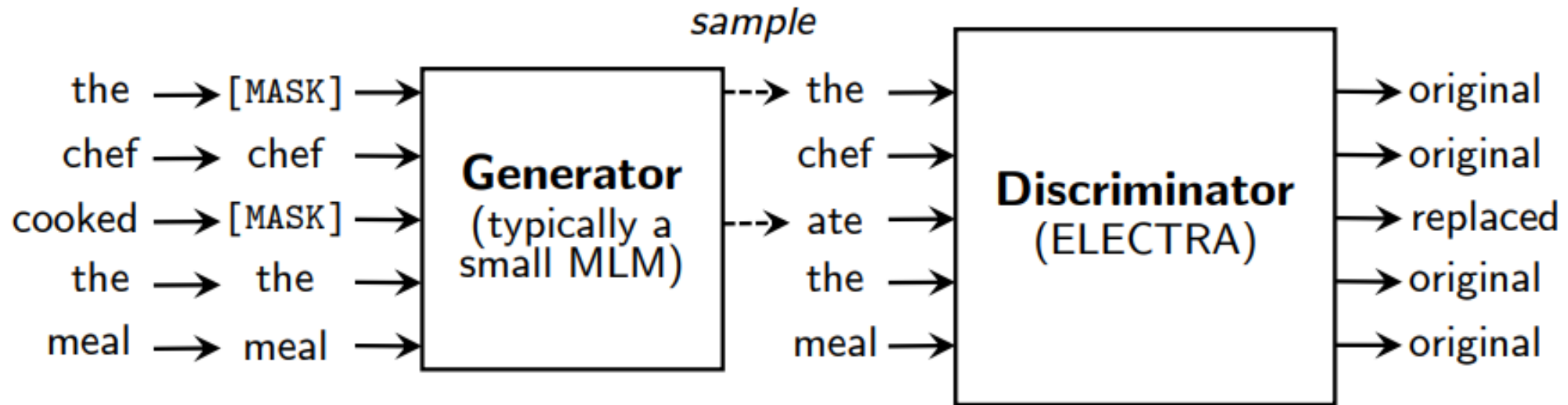
ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations.(Google Research, Toyota Technological Institute at Chicago)

| | |
|----|--|
| 목적 | 큰 네트워크 훈련 시 성능 향상을 기대하지만 실제 Large network를 train 할 때 메모리 및 속도 문제 발생 |
| 방법 | Parameter reduction technic 1. Factorized embedding parameterization : 기존 BERT는 input embedding vector size(context 학습 X)와 hidden layer output embedding vector size(context 학습 O)가 동일 → 굳이 size를 맞추는 필요가 없음 : 따라서 두 개의 작은 매트릭스로 나누어 파라미터 수를 줄일 수 있음 2. Cross-layer parameter sharing : parameter sharing 방법으로 depth에 따라 parameter가 커지는 것을 방지 SOP(Sentence-Order Prediction) |
| 코드 | https://github.com/google-research/ALBERT |

ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS. ELECTRA, 85, 90. (Stanford, Google)



ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning
Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2016). ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS. ELECTRA, 85, 90. (Stanford, Google)

| | |
|----|---|
| 목적 | BERT는 MLM을 통해 양방향으로 학습이 가능하지만 masking된 단어 하나만 예측하므로 지식 습득을 위해서는 다량의 코퍼스가 필요 |
| 방법 | <p>Replaced token detection task</p> <ul style="list-style-type: none">- Generator + Discriminator → GAN 구조를 따르지만 적대적 학습은 하지 않음 : generator를 이용하여 실제 입력의 일부 토큰을 그럴싸한 가짜 토큰으로 바꾸고, discriminator로 각 토큰을 조사하여 실제/가짜(생성된) 를 예측- Generator: 토큰을 masking 하는 대신 token을 적절한 대안으로 대체하여 input 수정하며 pre-training 이후에는 generator 제거- Discriminator: token ID를 예측하는 대신 각 token이 generator로 예측되었는지 여부 예측 |
| 코드 | https://github.com/google-research/electra |

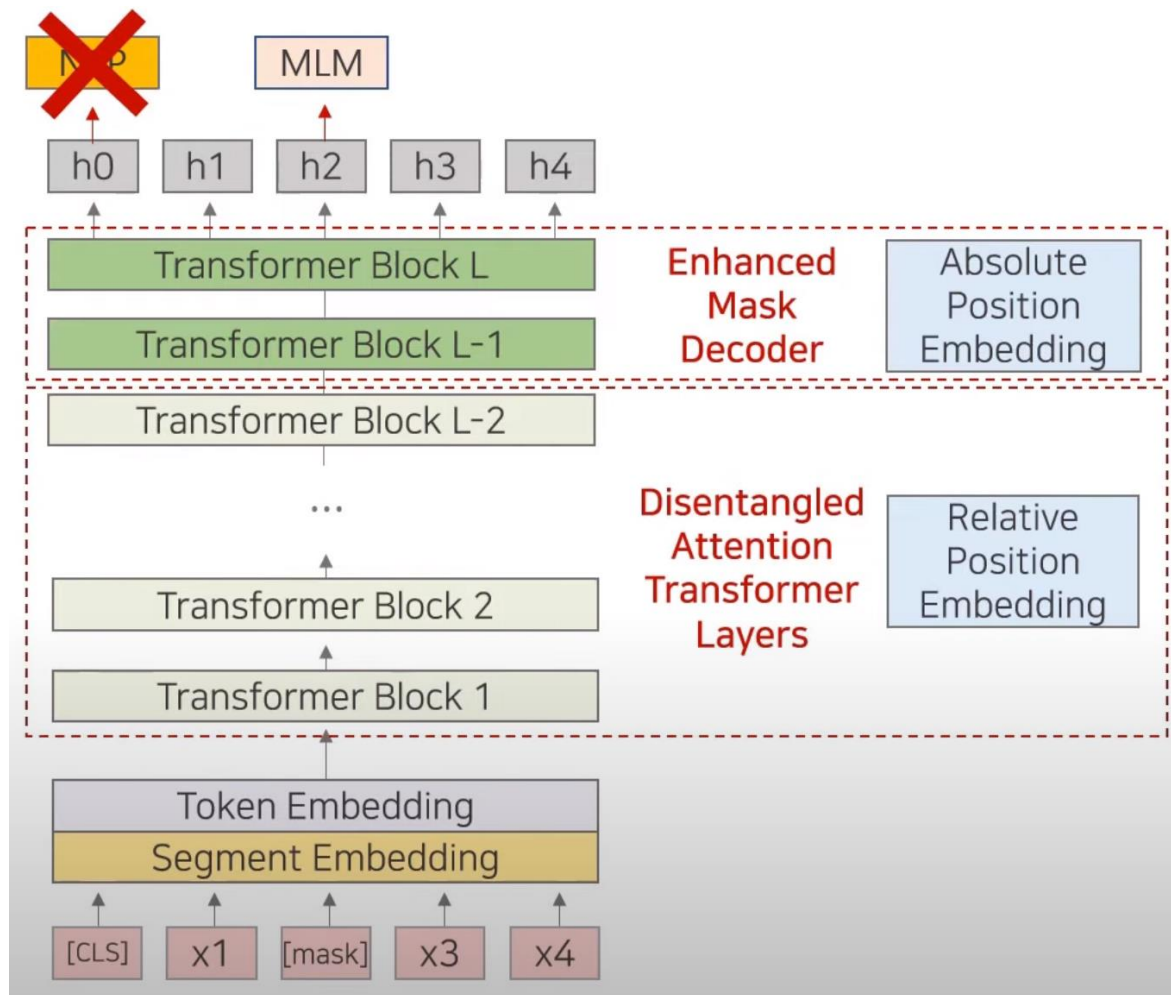
GPT3: Language Models Are Few-Shot Learners

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. (OpenAI)

| | |
|-------|---|
| 목적 | <ol style="list-style-type: none">1. Task에 따라 매번 fine-tuning 필요2. GPT-2에 task를 입력하여 범용성을 증가시키려 하였으나 fine-tuning을 거친 모델에 비해 성능이 다소 떨어짐 |
| 방법 | <p>GPT-3</p> <ol style="list-style-type: none">1. GPT-2 구조와 동일하나 GPT-2에서 모델의 크기 데이터셋의 크기 및 학습 횟수를 전반적으로 늘린 모델2. 1750억 개의 파라미터 학습 |
| 장점 | <ol style="list-style-type: none">1. 범용성 증가 |
| 단점 | <ol style="list-style-type: none">1. 여전히 양방향으로 정보를 학습하지 못하기 때문에 문맥 이해도는 낮음 → 다음 단어/문장 예측에는 강세를 보이지만 빈칸 맞추기, 두 문단 비교하고 답하는 작업, 긴 문단을 읽고 짧은 답변을 생성하는 태스크에서는 잠재적으로 낮은 성능을 보여줌2. 워낙 큰 모델이기 때문에 시간 및 비용이 많이 듦 |
| 예시 샘플 | https://ggoorr.net/thisthat/14882950 |

DeBERTa: Decoding-enhanced BERT with Disentangled Attention

He, P., Liu, X., Gao, J., & Chen, W. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In International Conference on Learning Representations., 2021.(Microsoft Dynamics 365 AI, Microsoft Research)



DeBERTa: Decoding-enhanced BERT with Disentangled Attention

He, P., Liu, X., Gao, J., & Chen, W. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In International Conference on Learning Representations., 2021.(Microsoft Dynamics 365 AI, Microsoft Research)

1. Disentangled representation

Disentangled: 얽혀 있는 것을 풀다

- 현재 embedding vector = content(단어 정보) + position(상대적인 위치 정보)
→ 이 두 정보를 분리하여 볼 필요가 있음
- 예) 'deep learning'이 같이 붙어 있을 때 강한 의미를 갖는다고 학습해야 함
→ Learning deep이 강한 의미를 갖는다고 학습하면 안됨. 따라서 relative position을 분류하여 학습 필요
- 기존 attention과 다르게 content-to-content, content-to-position, position-to-content 세 개의 attention 계산
 - Content-to-content: 단어와 단어 사이 attention
 - Content-to-position: 내가 궁금한 단어의 위치가 정해졌을 때, 다른 단어들은 어떤 상대적 위치를 갖는가?
 - Position-to-content: 내가 궁금한 단어의 상대적 위치가 정해졌을 때, 나랑 관련있는 단어는 무엇인가?

DeBERTa: Decoding-enhanced BERT with Disentangled Attention

He, P., Liu, X., Gao, J., & Chen, W. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In International Conference on Learning Representations., 2021.(Microsoft Dynamics 365 AI, Microsoft Research)

2. Enhanced mask decoder(EMD)

- 현재는 상대적인 위치만 고려
- 그러나 실제로 문장에서 절대적인 위치는 중요함
한국어 기준 → 주어는 앞 부분, 목적어는 중간 부분, 동사는 끝 부분
- 따라서 절대적인 위치 정보를 무시할 수 없으므로 마지막에 예측하기 전에 absolute position embedding 정보를 output에 더하여 예측 layer를 거치도록 유도

DeBERTa: Decoding-enhanced BERT with Disentangled Attention

He, P., Liu, X., Gao, J., & Chen, W. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In International Conference on Learning Representations., 2021.(Microsoft Dynamics 365 AI, Microsoft Research)

| | |
|----|---|
| 목적 | BERT 및 RoBERTa 모델을 개선하는 새로운 모델 아키텍처 생성하여 사전 훈련 단계의 효율 향상 |
| 방법 | <div>1. Disentangled Attention Mechanism : word embedding vector와 position vector를 독립적으로 인코딩 : content-to-content, content-to-position, position-to-content총 세 개의 어텐션을 구함</div> <div>2. Enhanced Mask Decoder(마지막 transformer encoder layer를 decoder라고 표현) : [Mask] token을 예측하는 layer에서 absolute position 정보를 추가하여 문장의 역할(주어, 목적어 등)을 학습 유도</div> <div>3. RoBERTa 참고 논문이므로 NSP 제거</div> |
| 코드 | https://github.com/microsoft/DeBERTa |

News



Google아, 너희들은 GPT-3 없지? 별 거 없네~

vs



큰 거 온다...
GPT-3 대항마 LaMDA 기반으로 하는 Bard
보여줄게. 기대해.