

KUBIG 2023-1

겨울방학 자연어처리(NLP) 분반

WEEK 7



Part 1

복습과제 리뷰

- BERT

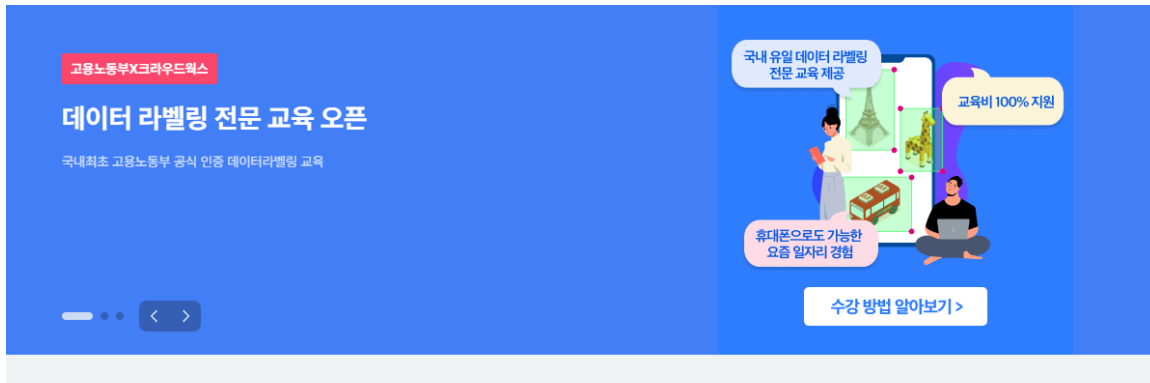


Part 2

7주차진도(1)

GPT

GPT



공지사항

[2023 국민내일배움카드] 최종평가 오답 해설 확인 안내

[2023 국민내일배움카드] 데이터라벨링 과정 안내

실습 프로젝트 안내(반려 및 정답보기, 진도를 관련)

데이터 라벨링이란?

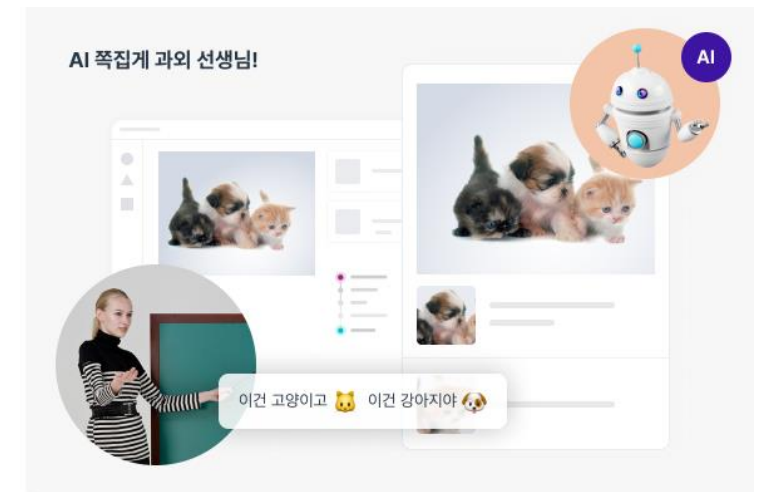
데이터 라벨링이란, 인공지능이 특정 데이터를 스스로 학습할 수 있도록 데이터를 알맞게 수집 및 가공하는 작업을 말합니다.



2022년 고용노동부
직업사전 정식등재



4차 산업혁명 시대
떠오르는 직업



GPT

Supervised learning
-> labeled data

	document	label
	아 더빙.. 진짜 짜증나네요 목소리	0
	홀...포스터보고 초딩영화잖...오버연기조차 가법지 않구나	1
	너무재밌었다그래서보는것을추천한다	0
	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
	사이폰페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 뺏겨보이지만 했던 커스틴 ...	1

Q	A	label
1 좋아하는 남자 때문에 너무 기분이 안 좋아.	무슨 일이 있었나봐요.	2
2 존잘할 수 있는 사람 만나고 싶어.	그런 사람 만날 수 있을 거예요.	2
3 오늘 일글보고 확실히 찍어줬어	실지 않을 걸장이었을텐데 당고생 많았어요.	1
4 이별4주년	새로운 연애를 시작해보시는건 어때세요.	1
5 파디나 몰카	기본전환을 하보세요.	0
7 여자들은 좋아하는데도 답장시간이 느릴 수 있나요?	일단 물어보면 느려졌죠.	2
8 혼자 있고 싶다	혼자만 있지 마세요.	0
9 두 사람만 좋으면 되는줄 알았는데 아니었나봐	현실의 벽이 부딪혔나봐요.	0
10 점 빼려 가고 싶다.	백만 골목짜질 거예요.	0

Unsupervised learning
-> unlabeled data

관람객 3시간이 소중했다. 난 타이타닉 때문에 영화를 좋아하게 됐는지도 모른다. 언제 다시 개봉할지 난 잘 모르겠지만, 그 날이 꼭 올거라고 믿고 기다릴게. 그 때는 다른 사람과 함께 보았으면 좋겠다.

s2lo**** | 2018.02.02. 23:59 | 신고



댓글 1 고귀한 꽃님이 | 2023-02-22 20:43:41

인스타 팔로우 완료했습니다^^



익명
02/22 15:00

(홍보) 고려대 친구들과 선배들과 밥먹할 수 있는 앱!!



댓글 2 사나운 뱀속 | 2023-02-22 22:25:29

펼~~~~럭

안녕하세요! 저희는 고려대 20학번에 재학 중인 학생 두명입니다. 저희가 학교를 다닌지 무려 3년이 됐지만 여전히 중강시간에 밥 같이 먹을 친구를 찾기가 힘들고 귀찮아서 친구들과, 선배들과 밥먹할 수 있는 앱을 만들어 봤습니다!



댓글 3 고귀한 중계중 | 2023-02-22 22:26:48

그냥 계속 걸어놔으면

——— 앱 간략한 소개 ———
도보 30분 이내에 있는 친구들과 일정 시간이 지나면 폭발하는 타임밤 알람을 만들어 서로의 상태를 확인하고 편하게 밥먹, 술먹 등을 걸 수 있는 앱입니다!

——— 다운로드 링크! ———

🍏 [iOS 다운로드 링크]
<https://apps.apple.com/kr/app/timebomb/id1658754766?l=en>



댓글 4 슬픈 단로박 | 2023-02-22 22:45:25

이쁘다

GPT

Supervised learning
-> labeled data

- 수작업

- 많은 시간과 자원을 필요로 함

	document	label
	아 더빙.. 진짜 짜증나네요 목소리	0
	졸...포스터보고 초딩영화줄...오버연기조차 가엾지 않구나	1
	너무재밌었다그래서보는것을추천한다	0
	교도소 이야기구먼..솔직히 재미는 없다..궁정 조정	0
	사이폰페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 뿔어보이거만 했던 커스틴 ...	1

Q	A	label
1. 좋아하는 남자 때문에 너무 기분이 안 좋아.	무슨 일이 있었나요.	2
2. 혼잡할 수 있는 사람 만나고 싶어.	그런 사람 만날 수 있을 거예요.	2
3. 오늘 일종료고 확실하 제아껴	일치 않을 것일지언정원의 할고성 있었어요.	1
4. 이별4년	이별 4년 전까지 계속되었던 것이지요.	1
5. 파디나	파디나에 대해 말씀드릴게요.	0
6. 여자들만 웃는 그날도 엄청시간이 느릴 수 있네요!	당연할까봐 놀라겠죠.	2
7. 혼자 있고 싶다	혼자만 있지 마세요.	0
8. 두 사람이 좋으면 되는데 맞았는데 아니었나봐	현실의 벽이 부딪혔나봐요.	0
9. 잘 배려 하고 싶다.	배려 할것처럼 거예요.	0

Unsupervised learning
-> unlabeled data

- 전이 (transfer) 학습에 효과적인 목적함수를 정의하는 것이 불분명하다

- 학습된 표현을 다른 과제로 전이하는 가장 효과적인 방법에 대한 일치된 의견이 없다

3시간이 소중했다. 난 타이타닉 때문에 영화를 좋아하게 됐는지도 모른다. 언제 다시 개봉할지 난 잘 모르겠지만, 그 날이 꼭 올거라고 믿고 기다릴게. 그 때는 다른 사람과 함께 보았으면 좋겠다.

s2lo**** · 2018.02.02 23:59 · 신고

인스타 팔로우 완료했습니다^^

익명 · 02/02 15:00

(중복) 그러대 친구들과 선택들과 발악할 수 있는 열매!

안녕하세요! 저희는 고려대 20학번에 재학 중인 학생 두명입니다. 저희가 학교를 다닌지 무려 3년이 됐지만 여전히 공강시간에 밥 같이 먹을 친구를 찾기가 힘들고 귀찮아서 친구들과 선후배들과 발악할 수 있는 열매를 만들어 왔습니다!

댓글 2 사나는 행복 · 2023-12-22 22:55:29

필~~~~력

댓글 4 슬픈 단로박 · 2023-12-22 22:45:25

이쁘다

[iOS 다운로드 링크]
https://apps.apple.com/kr/app/timebomb/id1658754766?l=en

GPT

Generative Pre Training of a Language Model

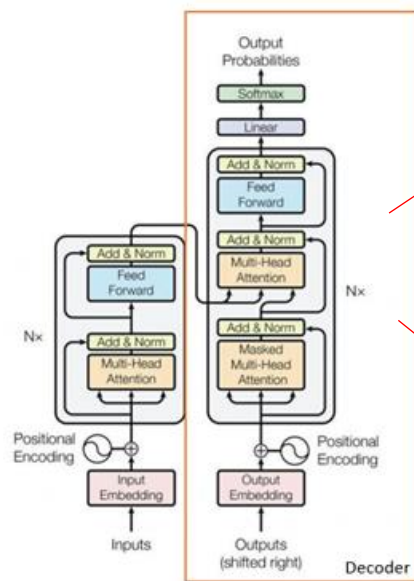
[unsupervised
pre-training]

+

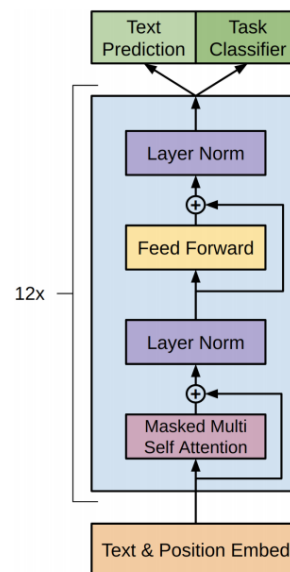
supervised
fine-tuning]

: 대량의 unlabel text 데이터를 통해 높은 수준의 언어모델을 학습

: label된 분류 데이터를 통해 특정 과제에 맞춰 fine tuning



: Transformer



GPT

I Generative Pre Training of a Language Model

[**unsupervised
pre-training**

+

**supervised
fine-tuning**]

: 대량의 unlabel text 데이터를 통해 높은 수준의 언어모델을 학습

: label된 분류 데이터를 통해 특정 과제에 맞춰 fine tuning

$$L_1(\mathcal{U}) = \sum \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta)$$

$$h_0 = UW_e + W_p$$

$$h_l = \text{transformer_block}(h_{l-1}) \forall i \in [1, n]$$

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

▶ 학습 시 i 시점 이후의 값은 masking -> AutoRegressive

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

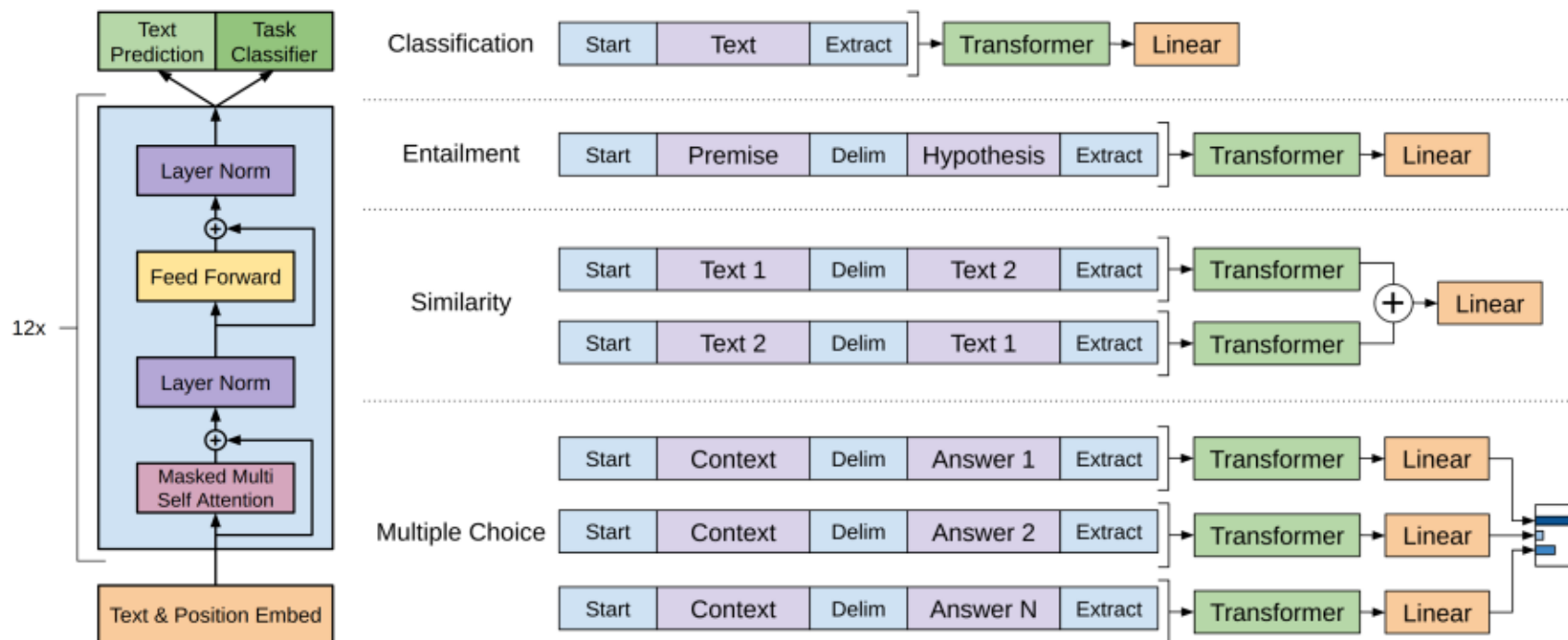
$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y | x^1, \dots, x^m).$$

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

GPT

Generative Pre Training of a Language Model

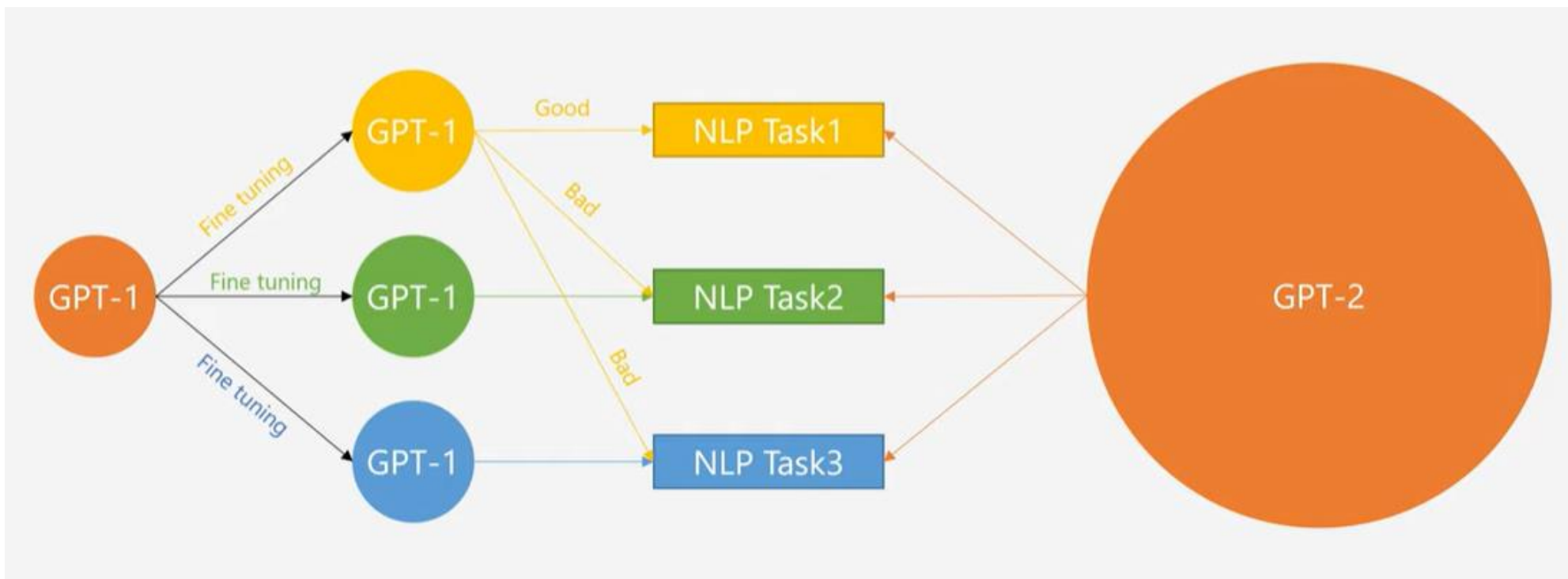
: Task-specific input transformation => no task-specific extra layer



분류 이외의 task에선 두 개 이상의 문장을 구분하기 위한 special token Delim (\$)을 활용

GPT-2

I Language Models are Unsupervised Multitask Learners (Radford 2019)



fine-tuning과 같이 지도학습을 활용한 좁은 범위의 문제만을 해결하는 것이 아니라
데이터를 수동으로 분류하는 과정 없이도 **더 범용적인 모델**을 제시

GPT-2

I Approach

$p(\text{output} \mid \text{input})$

$$p(x) = \prod_{i=1}^n p(s_i \mid s_1, \dots, s_{i-1})$$



언어모델 구조가
여러 다른 task까지도
다 다룰 수 있도록

$p(\text{output} \mid \text{input, task})$

Translation : (*translate to french, english text, french text*)
QA : (*answer the question, document, question, answer*)

“I hate the word ‘**perfume**,’” Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

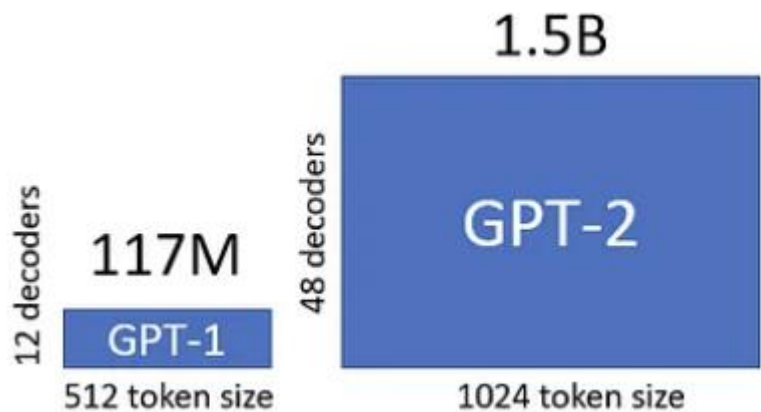
If listened carefully at 29:55, a conversation can be heard between two guys in French: “-**Comment on fait pour aller de l’autre côté? -Quel autre côté?**”, which means “- **How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.



GPT-2



Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

Bigger model

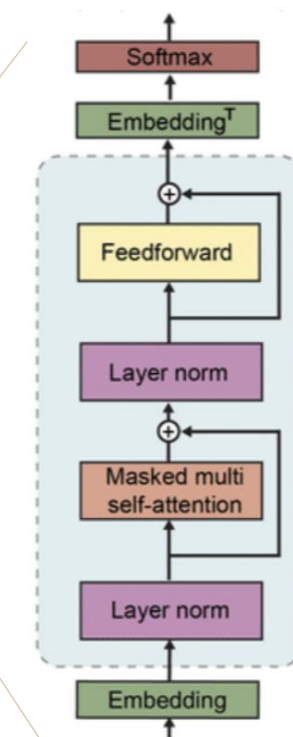
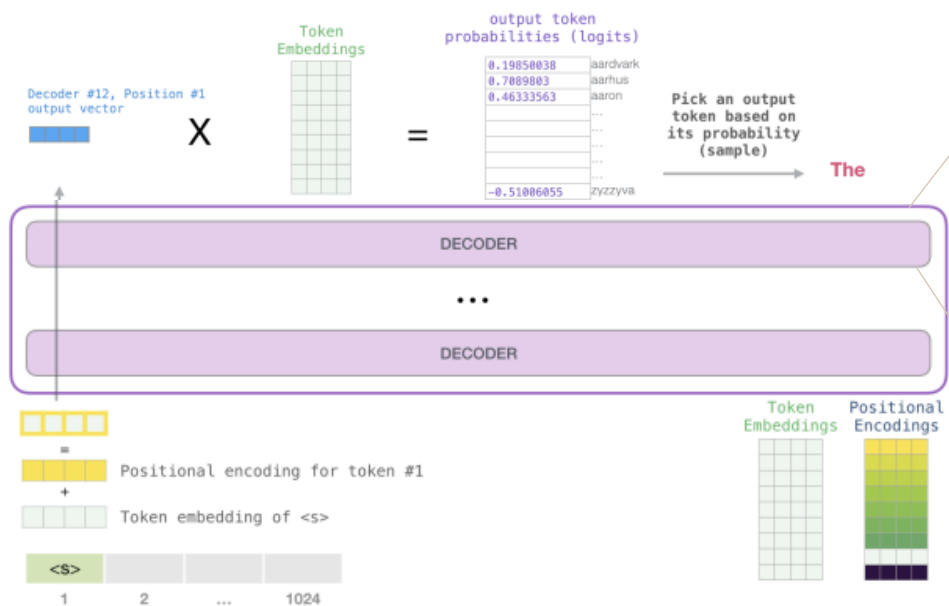


Web Scape from reddit
40GB text
Curated/filtered by humans
(WebText)

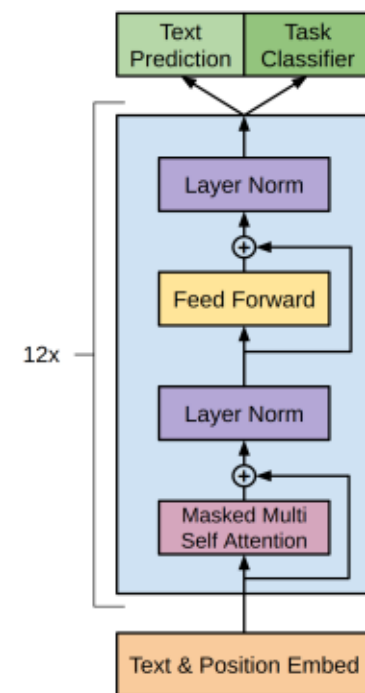
Bigger Dataset

GPT-2


Architecture



GPT-2 decoder block



GPT-1 decoder block

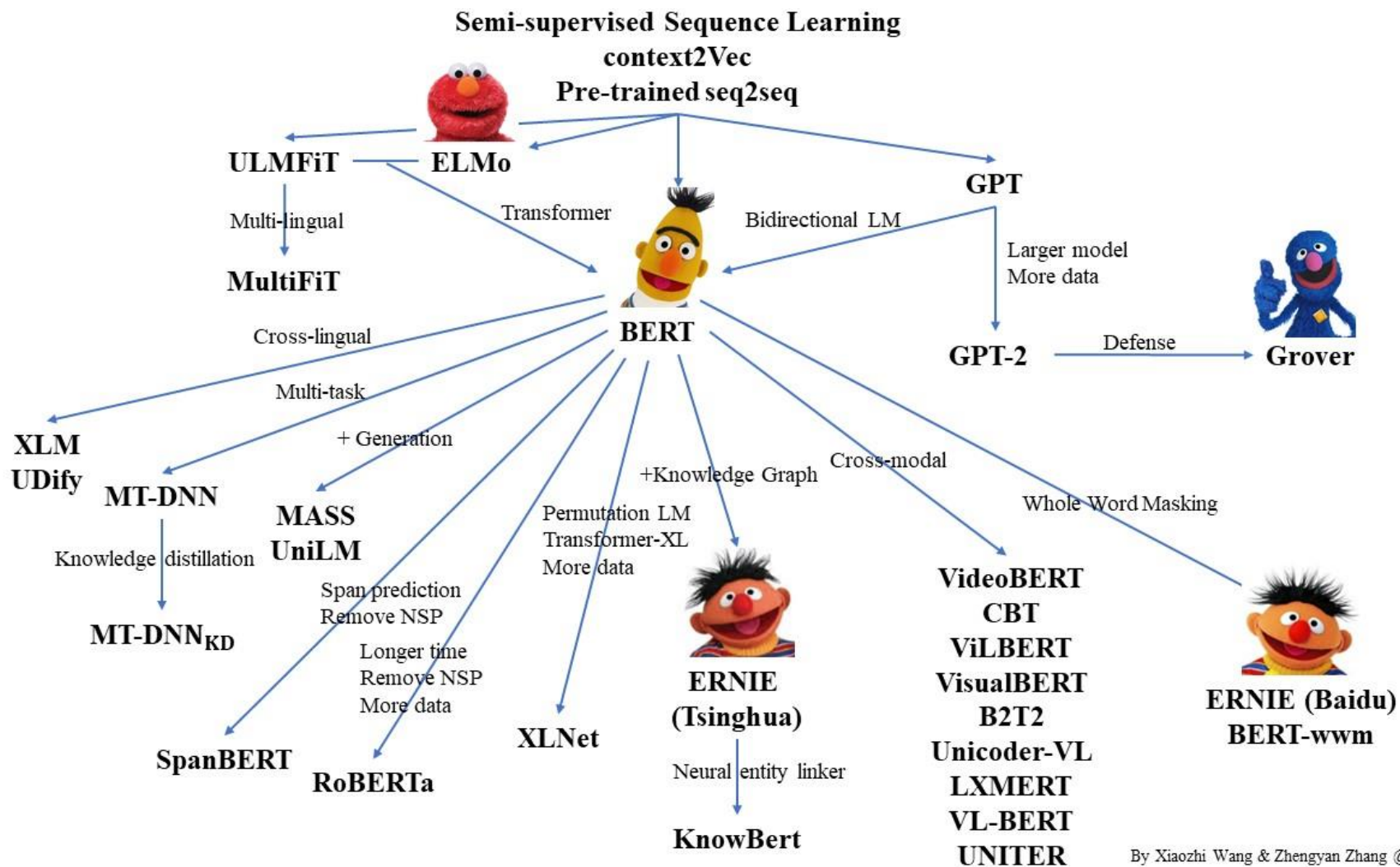


Part 3

7주차진도 (2)

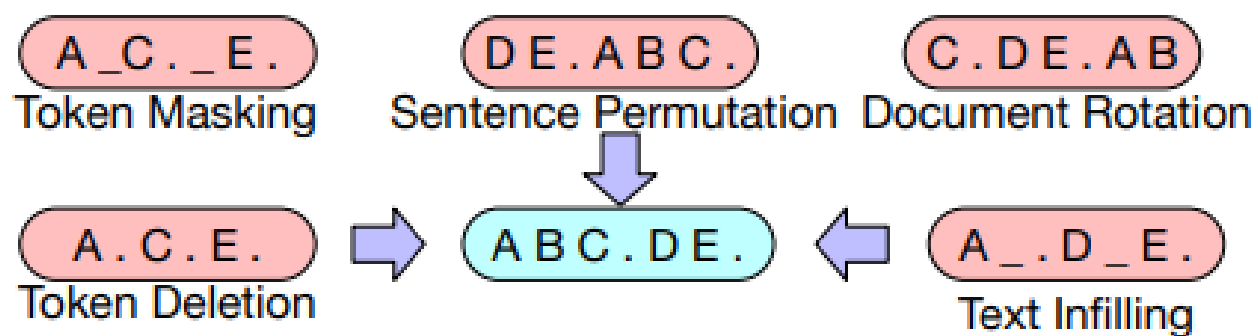
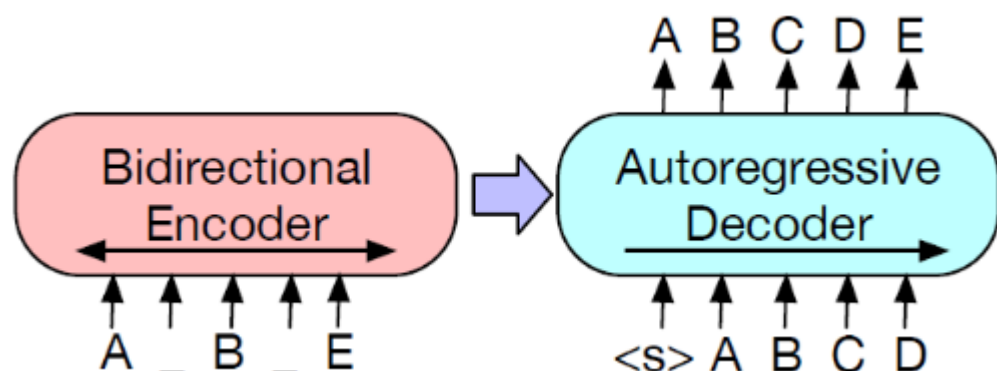
BERT Variants

Beyond BERT

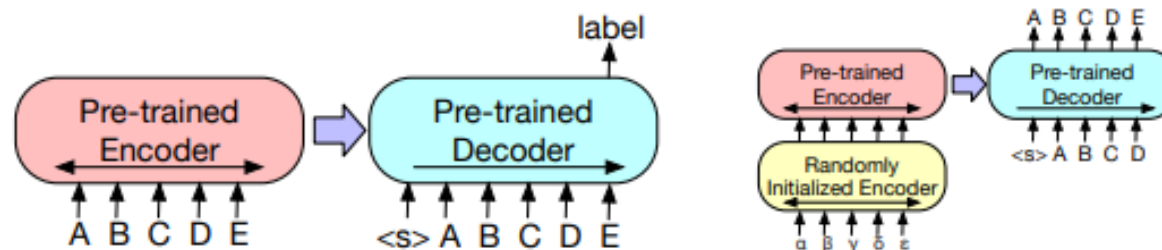


By Xiaozhi Wang & Zhengyan Zhang @THUNLP

BART



- AE + AR : Transformer 구조와 유사, ReLu -> GeLU
- Noise flexibility : corrupted된 데이터를 bidirectional encoder에 집어넣어 decoder가 복원시키게끔 학습됨
(fine tuning시에는 noise X)





Part 4

예습과제리뷰

- kogpt2 문장생성

```
[ ] sent = '문재인 대통령은 정책 발표를 위해'
    input_ids = tokenizer.encode(sent)
```

```
[ ] import random

while len(input_ids) < 50:
    output = model(np.array([input_ids]))
    top5 = tf.math.top_k(output.logits[0, -1], k=5)
    token_id = random.choice(top5.indices.numpy())
    input_ids.append(token_id)
```

```
[ ] tokenizer.decode(input_ids)
```

'문재인 대통령은 정책 발표를 위해' 지난 18일부터 22일 사이에 청문회와 국회 국정감사를 잇따라 받았다. 청문회는 이달 30일 종료되며 국회는 다음 날인 29일까지 청문회 준비를 끝냈다. 여야 모두 국정운영 전반에 관한 청문을 통해

How to generate text: using different decoding methods for language generation with Transformers

Published March, 2020.

[Update on GitHub](#)



[patrickvonplaten](#)
[Patrick von Platen](#)

[Open in Colab](#)

```
import torch
from transformers import PreTrainedTokenizerFast
from transformers import GPT2LMHeadModel
```

```
tokenizer = PreTrainedTokenizerFast.from_pretrained('skt/kogpt2-base-v2')
model = GPT2LMHeadModel.from_pretrained('skt/kogpt2-base-v2')
```

The tokenizer class you load from this checkpoint is not the same type as
The tokenizer class you load from this checkpoint is 'GPT2Tokenizer'.
The class this function is called from is 'PreTrainedTokenizerFast'.

#tensorflow가 pytorch보다 런타임이 좀 더 길었다.

```
[ ] tokenizer.decode(output_ids)
```

‘서울에서 부산가는 가장 빠른 방법은 바로 지하철을 타고 가는 것이다.¶이렇게 되면 서울역까지 30분 정도 걸린다.¶그런데 이게 너무 길어서 승객들이 불편을 겪는다.¶그래서 대중교통전용지구를 만들기로 했다.¶부산시는 지난해 12월, 교통혼잡과 도시미관을 해치는 버스 노선을 없애고, 시내버스 노선을 대폭 축소하는 내용의 ‘대중교통의 육성 및 지원에 관한 법률’ 개정안을 입법예고했다.¶개정안은 현재 운행 중인 버스의 경우 정류장마다 요금을 부과하고 있다.¶하지만 앞으로는 모든 노선의 요금도 함께 내야 한다.¶또한 기존에는 일반버스와 마을노선만 적용됐지만’

학습되는 데이터셋이 매우 중요해보인다.

```
▶ sent = '배가 고프면'
input_ids = tokenizer.encode(sent)
```

```
[ ] import random
import numpy as np

while len(input_ids) < 50:
    output = model(np.array([input_ids]))
    top5 = tf.math.top_k(output.logits[0, -1], k=5)
    token_id = random.choice(top5.indices.numpy())
    input_ids.append(token_id)
```

```
[ ] tokenizer.decode(input_ids)
```

‘배가 고프면 안 되는 건지?¶내가 <unk>가 이거 왜 안 돼? 아.¶내가 안 되겠어요?¶이거 몰로 된거야. 넌 안 될 것 같은데?¶이제부터 내가 안 돼’

?ㅋㅋㅋㅋㅋ원소리야

https://huggingface.co/docs/transformers/internal/generation_utils

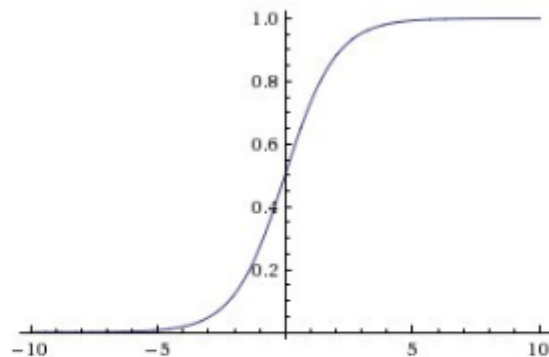
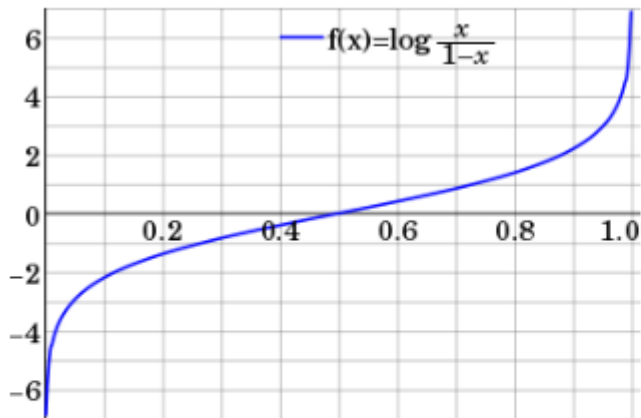
The output of `generate()` is an **instance** of a subclass of `ModelOutput`. This output is a data structure containing all the information returned by `generate()`, but that can also be used as tuple or dictionary.

`repetition_penalty` : The parameter for repetition penalty. 1.0 means no penalty.

<파라미터 설명>

`past_key_values` : contains precomputed key, value, hidden states of the attention blocks. Can be used to speed up decoding.

`use_cache` : if set `True`, `past_key_values` key value states are returned and can be used to speed up decoding (see `past`).



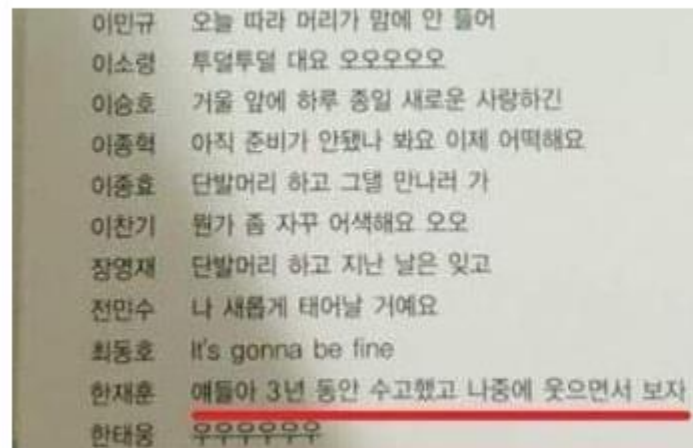
NN 마지막 layer에 적용되어 분류 문제를 해결
Network이 제시한 분류의 확률을 실수로 변환해줌

좌측은 logit. softmax 함수에 집어넣기 전 sigmoid의 x 상태(-inf inf)와 동일 / 우측은 sigmoid (0,1)

★ Logit function can transform those probabilities to real numbers.

<https://deeptai.org/machine-learning-glossary-and-terms/logit>

kakao brain의 kogpt(GPT3)를 사용하고 싶으나 kogpt는 무려 14GB의 램을 사용함 (Colab에서 기본 제공하는 ram의 크기는 12.7GB) 따라서 코랩에서는 결제하지 않는 한 kogpt 사용이 불가능한듯ㅜㅜ



이민규 오늘 따라 머리가 맘에 안 들어
이소령 투덜투덜 대요 오오오오오
이승호 거울 앞에 하루 종일 새로운 사랑하긴
이종혁 아직 준비가 안됐나 봐요 이제 어떡해요
이종호 단발머리 하고 그댈 만나러 가
이찬기 뭔가 좀 자꾸 어색해요 오오
장영재 단발머리 하고 지난 날은 잊고
전만수 나 새롭게 태어날 거예요
최동호 It's gonna be fine
한재훈 애들이 3년 동안 수고했고 나중에 웃으면서 보자
한태웅 우우우우우우

3학년 5반

김재상 잘 실렘



수고하셨습니다!

Contact

15기 김제성

✉ rlawptjd1409@korea.ac.kr

☎ 010-2609-5046