



Improving Language Understanding by Generative Pre-Training

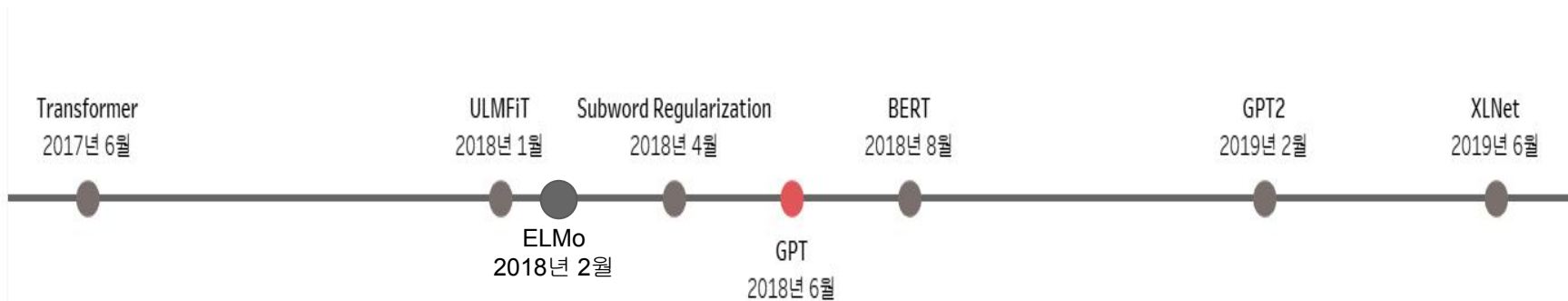
저자 : Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever (OpenAI)

발표자 : 박지민
(kpdpkp@gmail.com)

Why GPT?

Why GPT? 1) NLP에서의 Pre-Trained Model 설명

Pre-Trained Model Timeline



Why GPT? 2)Generative Model

Request a completion from the language model.



Brexit has already cost
the UK economy at
least £80bn since the
EU referendum, and
many industry experts
believe the economic
damage from Brexit will
be greater still



In some cases, the costs
will be far harsher. A
recent report published



기사 쓰는 인공지능 지피티

5월 11일 · 🌐

[인공지능 "지피티"가 쓴 가짜 뉴스입니다.]

제목: [AG] 한국, 사격 2연속 3관왕으로 아시안게임 최다 우승 타이

[마이데일리 = 중국 광저우 특별취재반] 한국 남자 사격이 2연속 3관왕에 오르며 아시안게임 사상 최다 메달과 금메달을 차지했다.

사격 남자 50m 소총복사 결선 경기가 펼쳐진 14일 오후 광저우 아오티사격장에서 한국은 김중현(25.창원시청)과 한진섭(29.충남체육회), 최용래(25.한화갤러리아)가 각각 금, 은, 동메달을 나눠 가졌다.

50m 소총복사 본선에서 1, 2위권을 유지하며 금메달이 유력해보였던 김중현은 결선에서 75점을 기록, 2위 한진섭(68/70.6점)과 동점을 이뤘지만 2발사결과 1.4점 뒤져 아쉽게 은메달에 머물렀다.

그러나 김중현은 지난 13일 남자 50m 소총복사 단체전에서 금메달을 획득한데 이어 이번 대회 2관왕에 올랐다.

지난 2006년 도하대회와 2008년 베이징올림픽에서 금메달을 목에 걸었던 김중현은 대회 2연패를 달성했고 한진섭은 개인 종목에서 2회 연속 금메달에 성공했다.

한국은 아시안게임 사격에서 2연속 3관왕을 달성한 선수가 등장하면서 중국의 거센 도전에 정면 대응할 수 있게 됐다. 또한 지난 1982년 뉴델리대회부터 총 20개의 금메달 중 12개를 차지한 한국은 3회 연속 중국에 이어 사격에서 최다 금메달과 최다 메달의 2개차 2위에 오르는 등 아시아 스포츠 최강을 지켜왔다.

한국의 뒤를 이어 북한은 여자 트랩의 송수주, 혼성 트랩의 이호림과 남자 권준철 등 2개 종목에서 2관왕을 차지했고 사격에서만 6개의 메달이 나왔다.

또한 사격대표팀은 지난 13일 개인전에서만 3관왕을 차지한 한국의 김준호와 함께 이번 대회 사격 종합 2위 수성에 앞장섰다.

[김중현, 사진 = 중국 광저우 XXX 기사 co.kr]기사 제목 및 보도자료 co.kr-NO.1 뉴미디어 실시간 뉴스 마이데일리(www.mydaily.co.kr)

TabNine

```
1 import os
2 import sys
3
4 # Count lines of code in the given directory, separated by file extension
5 def main(directory):
6     line_count = {}
7     for filename in os.listdir(directory):
8         _, ext = os.path.splitext(filename)
9         if ext not
```

not in line_count	Tab	15%
not in line_count.keys():	Tab+Tab	6%
not in line_count:	Tab+3	7%
not in line_	Tab+4	15%
not in	Tab+5	23%



Motivation

논문 목표

비지도학습으로 Unlabeled Text 활용해서 다양한 Specific Task를 풀고 싶음

“Our setup does not require these target tasks to be in the same domain as the unlabeled corpus.”

왜 해야하는가?

- 1) 도메인 데이터가 적음
- 2) 데이터를 구하려면 시간과 돈이 많이 들어감
- 3) Universal Representation을 학습

지도학습이 가능하더라도 비지도학습이 성능향상에 도움이 됨
ex) Word embedding

Challenging

1. 어떤 **Optimization Object** 가 효과적인지?

- **LM**, Translation, Discourse Coherence

2. **Transfer** 하는 방법에 대한 **Consensus**가 없음

- architecture를 task-specific하게 수정 (ELMo)
- **Task-aware input**사용 → Architecture 변화를 최소화
- Using intricate learning scheme(ULMFiT)
- **Adding auxiliary learning objective**

Framework

Generative
(Pre-Training)

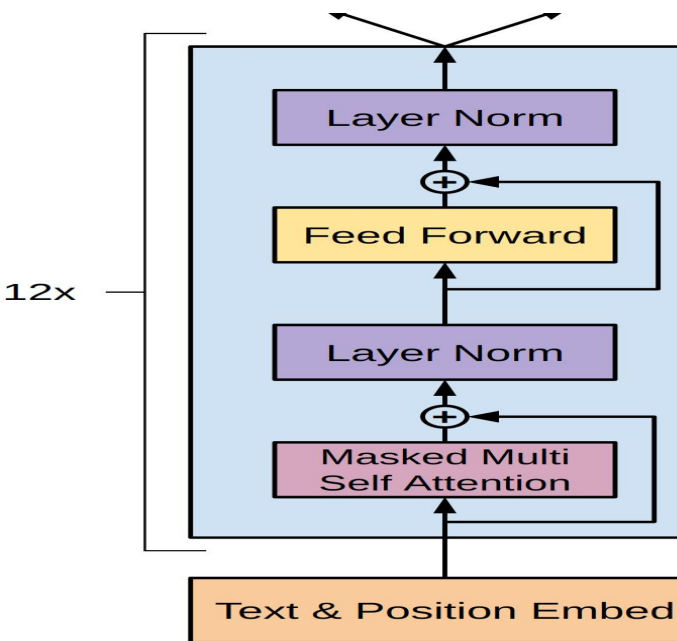
- Transformer Decoder
- Language Model



Discriminative
(Fine-Tuning)

- Input Transformation
- Auxiliary objective

Pre-Training 단계



Embedding

- BPE with 40,000 Merge
- Sinusoidal 대신 learned position embedding

Language Modeling

- Transformer Decoder
- BERT와 다른점

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

$$h_0 = UW_e + W_p$$

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

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

GPT vs Transformer

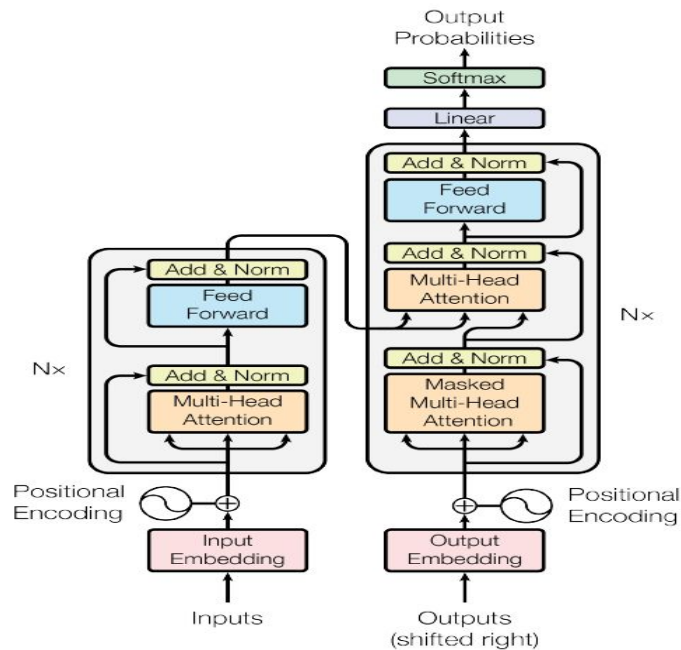
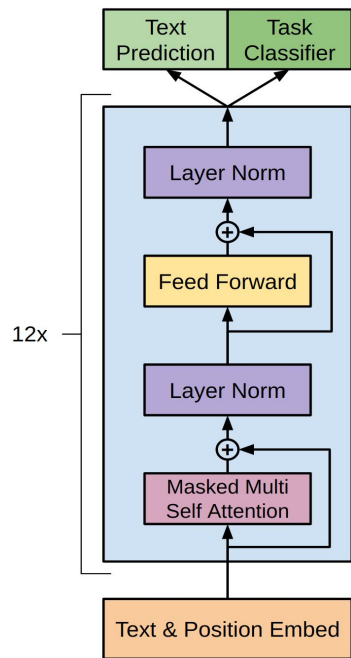
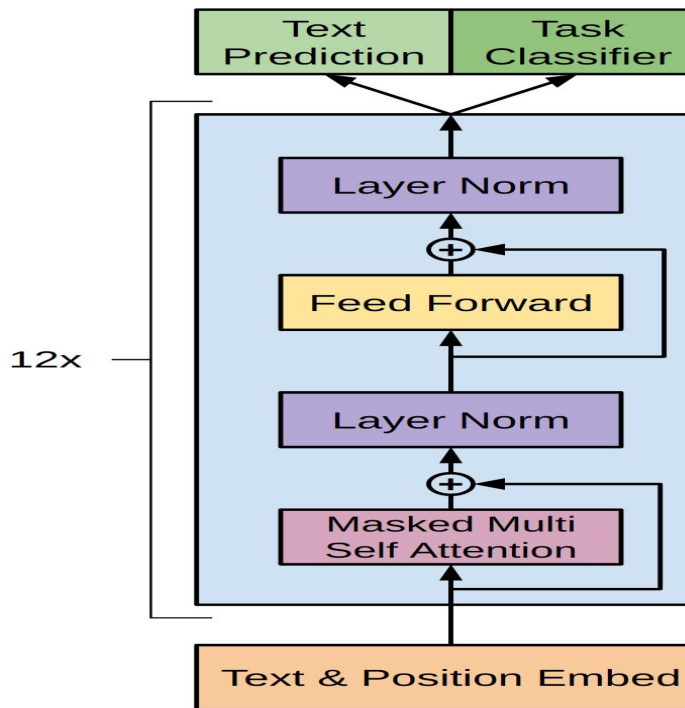


Figure 1: The Transformer - model architecture.

Fine-Tuning 단계



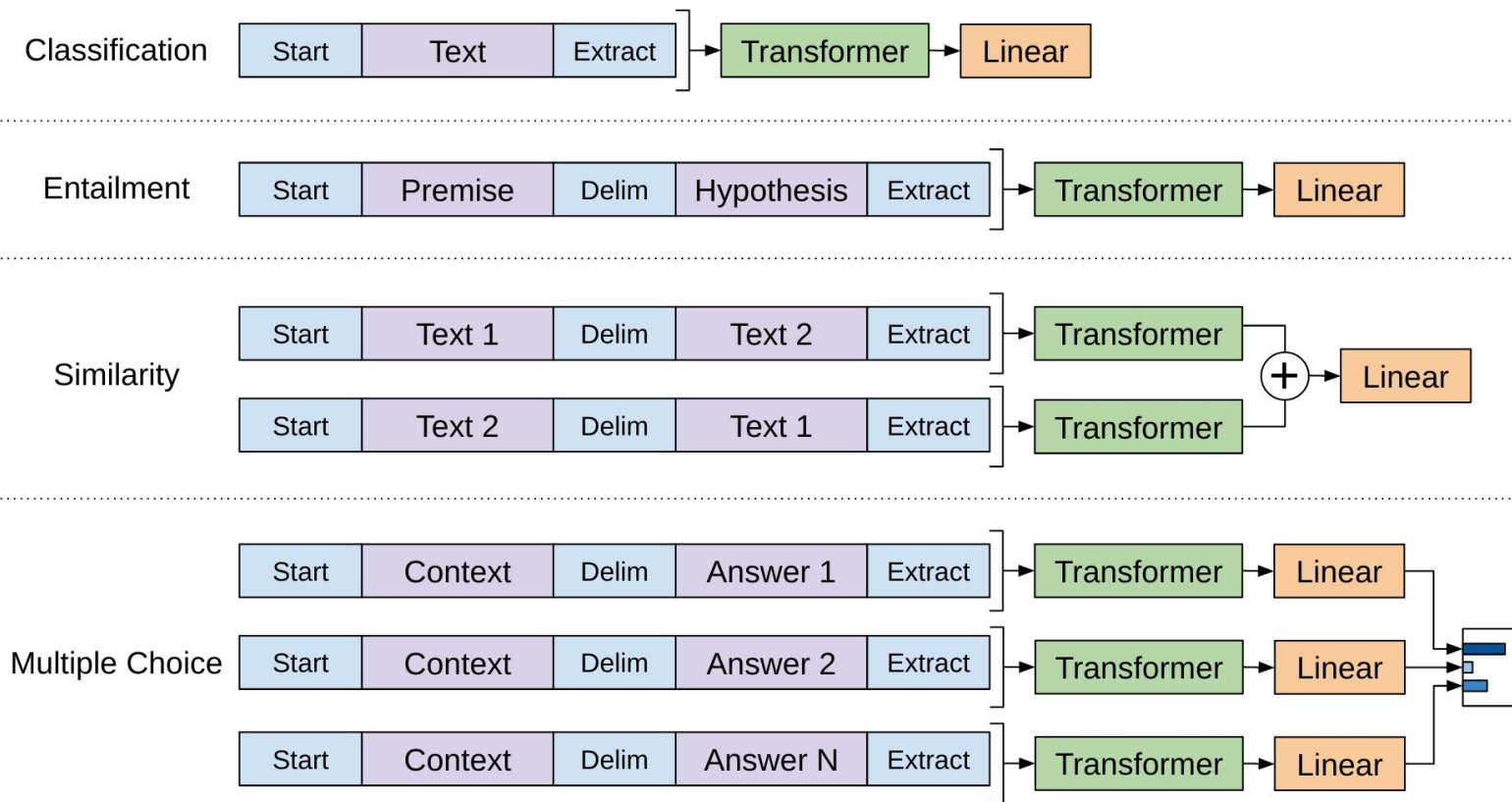
$$P(y|x^1, \dots, x^m) = \text{softmax}(h_l^m W_y).$$

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

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

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

Avoid changes to Model!



Textual Entailment

1. A black race car starts up in front of a crowd of people.
2. A man is driving down a lonely road.

Entail/ Contradict/ None

Story Cloze Test

ROCStories Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.

1. Karen became good friends with her roommate.
2. Karen hated her roommate.

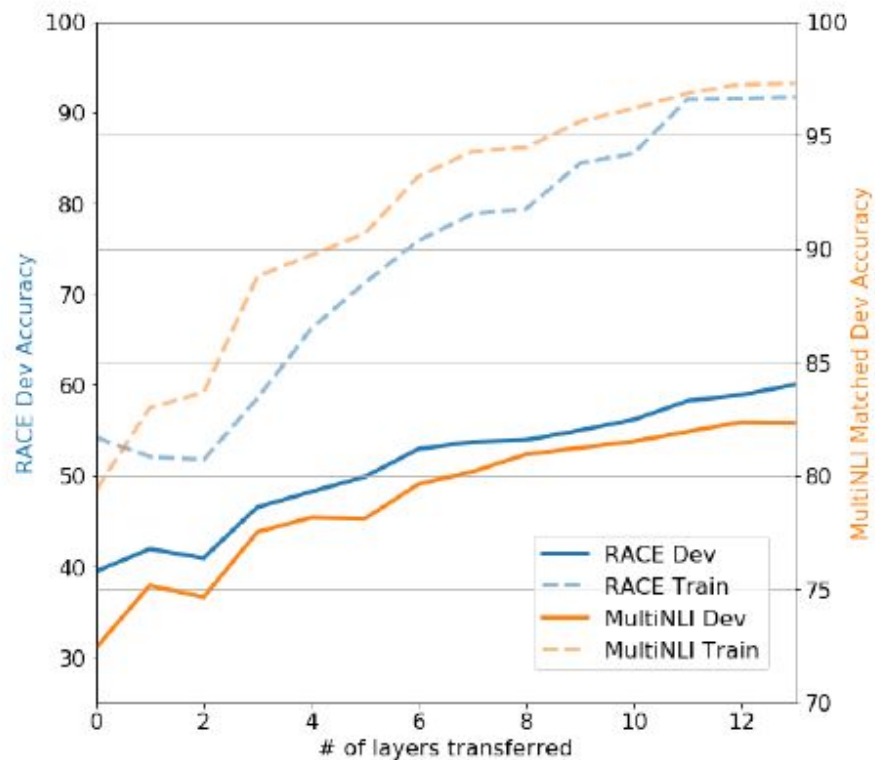
Performance

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>		
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	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

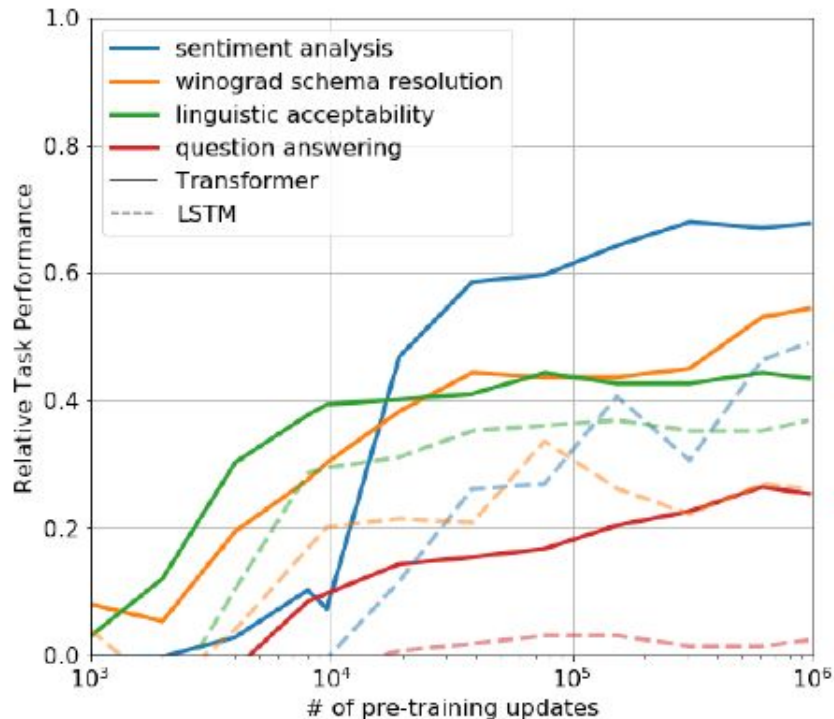
발표 당시 12개 dataset 중에 9개에서 sota 기록

- NLI에서 앞선 점수, 다른 task에서도 좋은 성적을 거두지 않았나하고 추측. 파고들지는 않았음.
- 엄청 작은 데이터셋(2490eg)인 RTE에서는 BiLSTM이 이김.
- Handle long-range contexts effectively.
- Corpus of Linguistic Acceptability(문법 판단, 언어 편향성 체크)에서 큰 성능 향상.
- 작은 데이터셋(5.7K)부터 큰 데이터셋(550K)까지 잘 학습함.

Analysis



Zero-shot Behavior (w/o fine-tuning)



왜 LM pre-training of transformer가 효과적?

Our hypothesis

- 1) the underlying generative model learns to perform many of the tasks we evaluate on in order to improve its language modeling capability
- 2) more structured attentional memory of the transformer assists in transfer compared to LSTMs

감성 분석의 경우 모든 예제에 very라는 토큰을 넣고 LM의 output을 positive와 negative라는 단어만 나오게 함.

Ablation Studies

Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

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

- 1) Larger Dataset이 Auxiliary Objective 효과가 큼.
- 2) LSTM보다 좋음.
- 3) Pre-Trained 안하면 성능이 대폭 하락.

Transformer가 나온지 꽤 됐지만 잘 활용하기가 쉽지 않았었나 봄. (뇌피셜)

FYI



François Chollet  @fchollet · 2월 18일

We all want safe, responsible AI research. The first step is not misrepresenting the significance of your results to the public, not obfuscating your methods, & not spoon-feeding fear-mongering press releases to the media.

That's our 1st responsibility.

- OpenAI는 GPT2가 악용될 소지를 걱정하여 모델을 제한하여 공개
- 적절한 조치였는가? (대중에게 공포심을 심어주는 것 아닐까?)