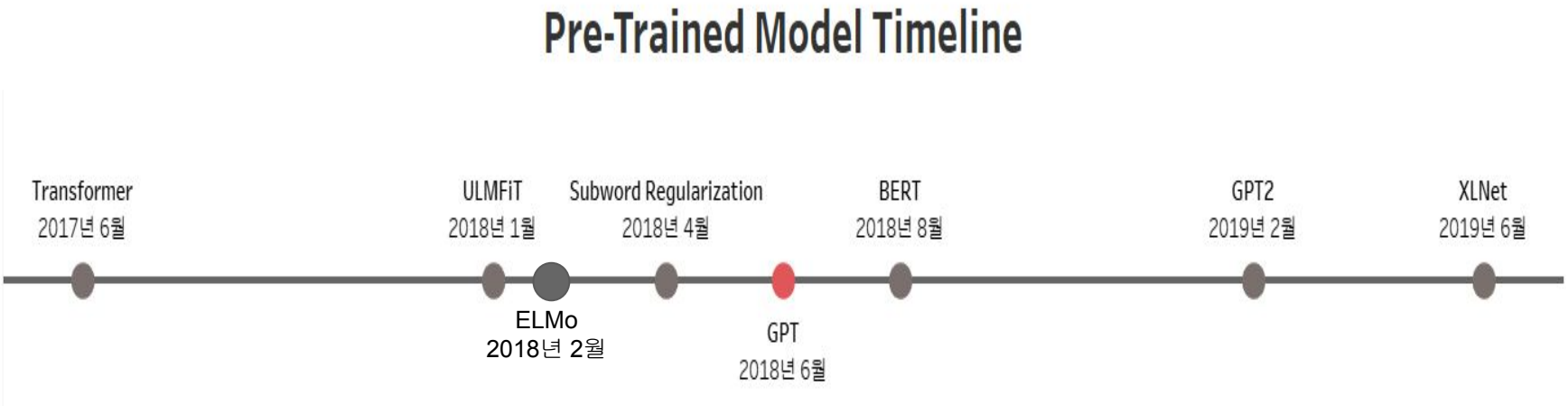
Improving Language Understanding by Generative Pre-Training

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

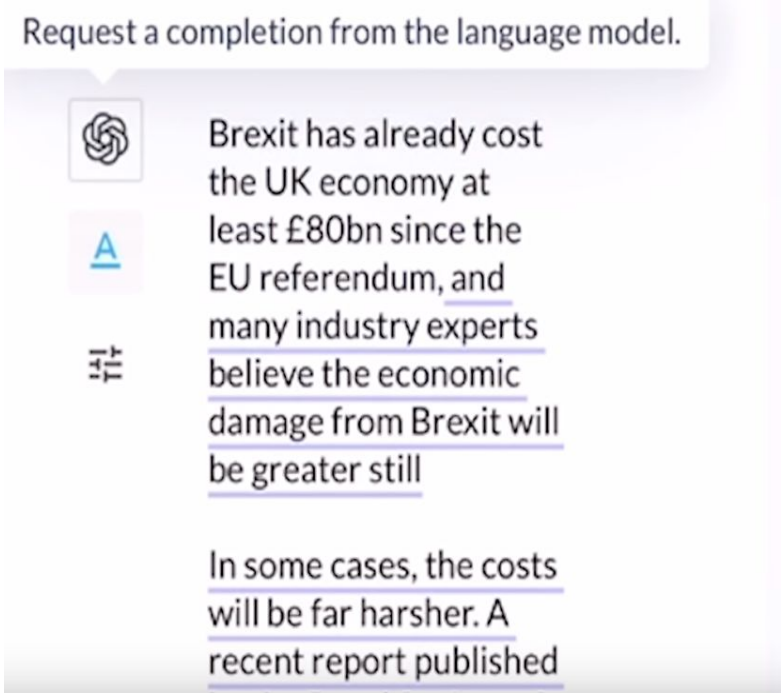
발표자 : 박지민
(kpdpkp@gmail.com)

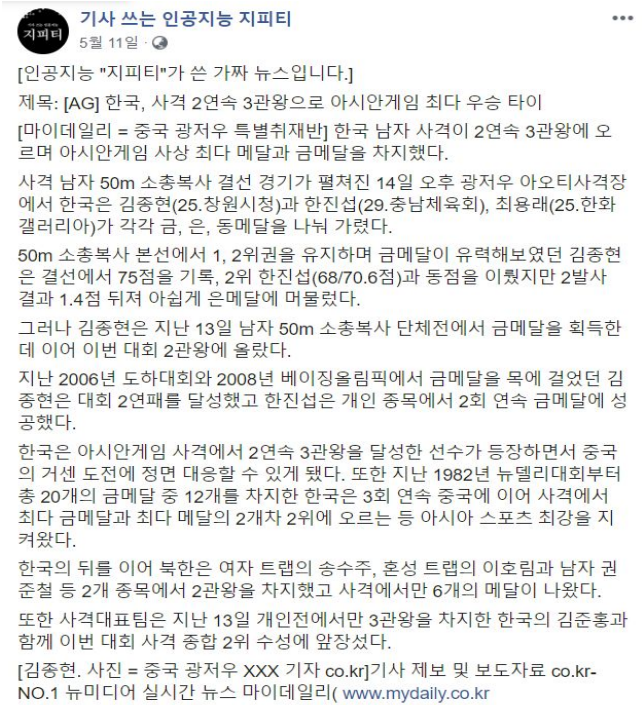
Why GPT?

Why GPT? 1)NLP에서의 Pre-Trained Model 설명



Why GPT? 2)Generative Model





영상링크 : https://www.youtube.com/watch?v=XMJ8VxgUzTc&fbclid=IwAR2xgRshy2wu0nnHXGy-pFOeooiXNPUMz1FViTBaeBh13zks5aQw3mOdDmo
기사쓰는 인공지능 지피티 : https://www.facebook.com/GPTisAI/?eid=ARBwuiCcaa0rXwHUEXnXizCAwfG18ydgmjND-4x6nFJVluhxkwcg4bxbS1yn4yY8gXVgANwluth6oqNm

TabNine

```
1  import os
2  import sys
3
4  # Count lines of code in the given directory, separated by file ex
5  def main(directory):
6      line_count = {}
7      for filename in os.listdir(directory):
8          _, ext = os.path.splitext(filename)
9          if ext not
10                 not in line_count                  Tab   15%
11                 not in line_count.keys():  Tab+Tab    6%
12                 not in line_count:           Tab+3    7%
13                 not in line_                 Tab+4   15%
14                 not in                       Tab+5   23%
15
16
17
```

https://www.youtube.com/watch?v=IssDX_zOwOo

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 input사용 → Architecture 변화를 최소화
   - Using intricate learning scheme(ULMFiT)
   - Adding auxilary learning objective

Framework

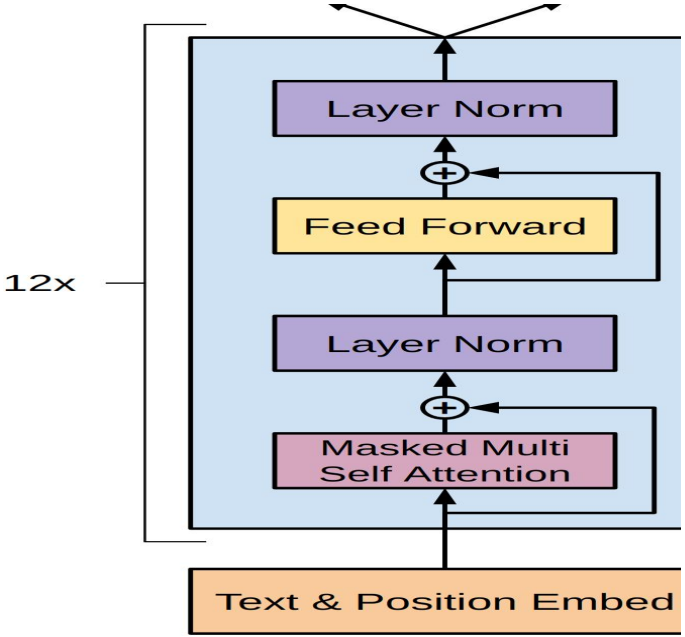
Generative
(Pre-Training)

- Transformer Decoder
  - Language Model

Discrimitive
(Fine-Tuning)

- Input Transformation
  - Auxilary objective

Pre-Training 단계



Embedding

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

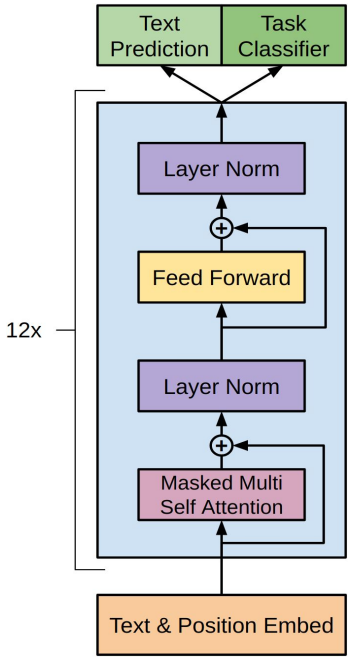
Language Modeling

- Transformer Decoder
- BERT와 다른점

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

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

GPT vs Transformer



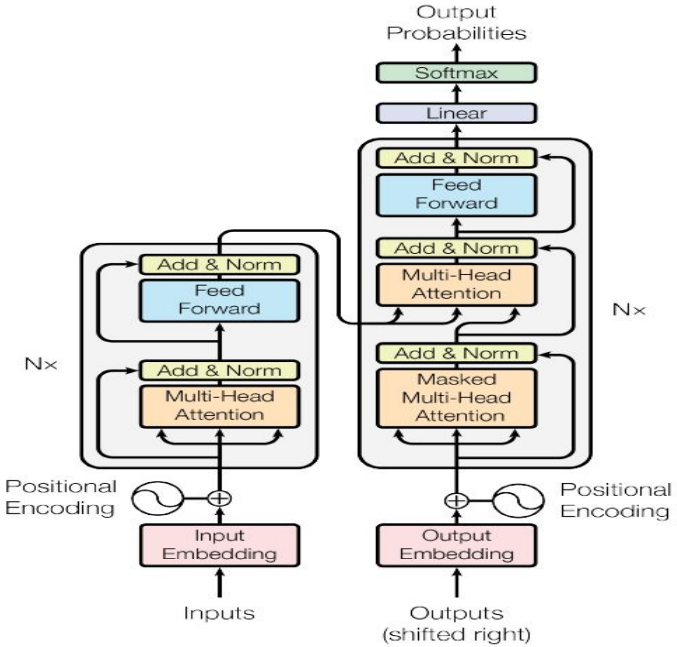
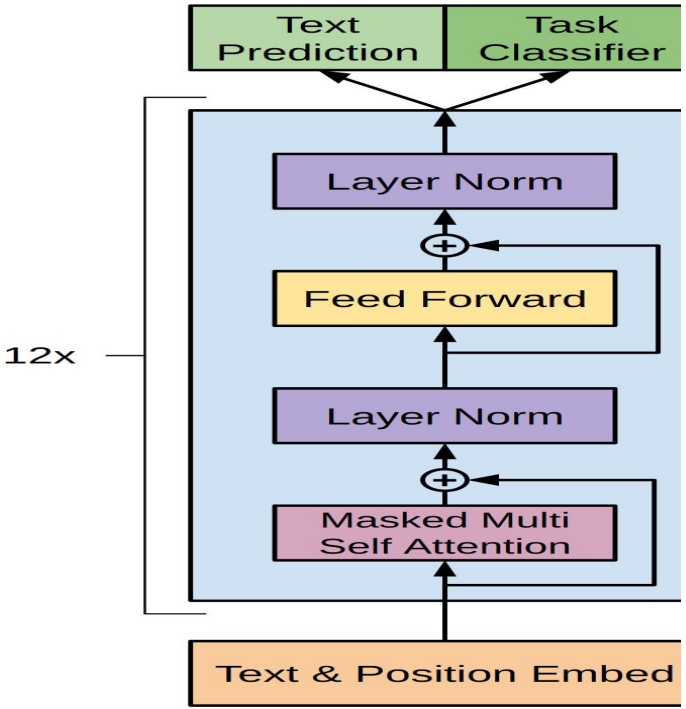


Figure 1: The Transformer - model architecture.

Fine-Tuning 단계



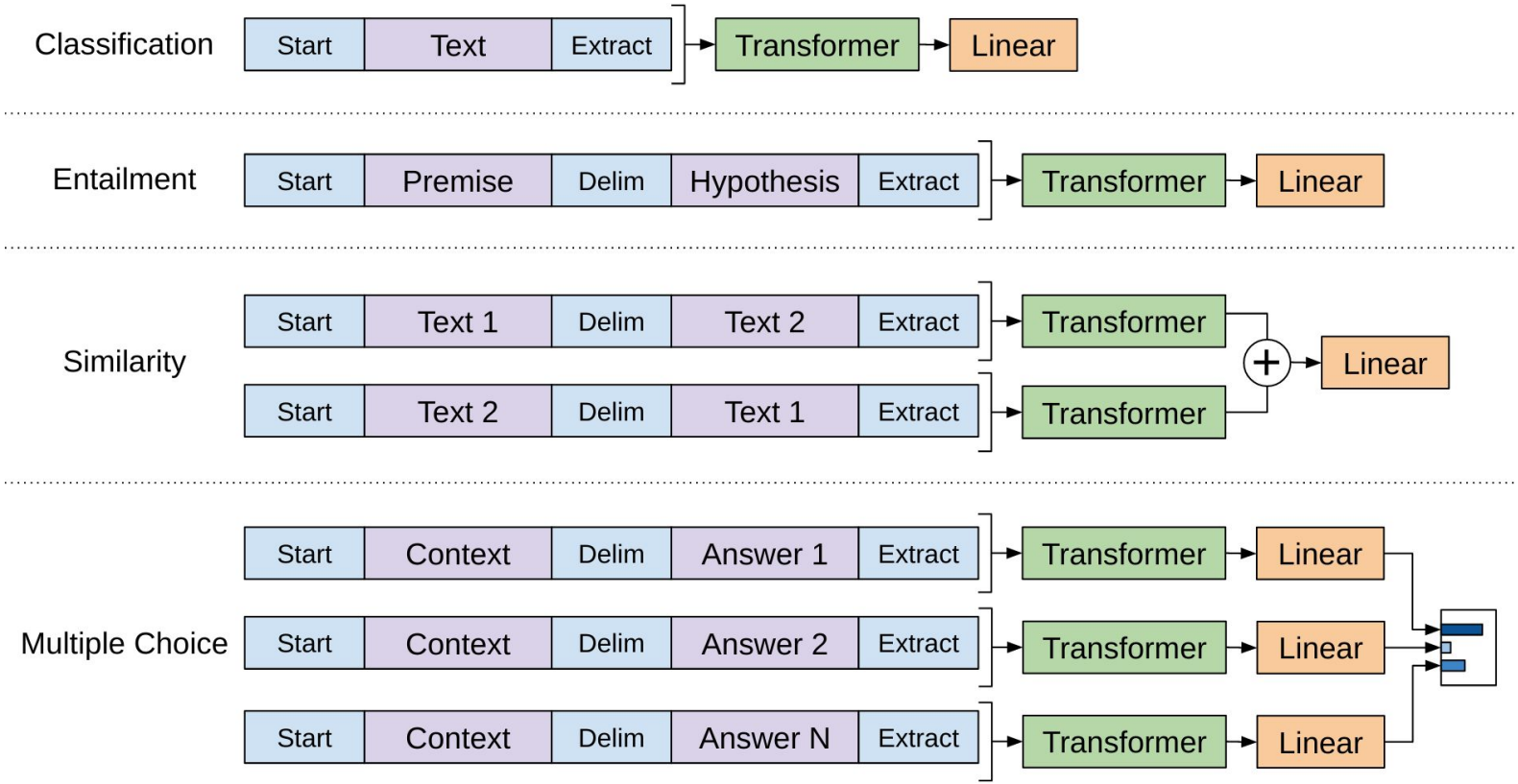
$$P(y|x^1,\ldots,x^m) = \texttt{softmax}(h_l^m W_y).$$

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

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

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

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

https://openai.com/blog/language-unsupervised/

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

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

https://openai.com/blog/language-unsupervised/

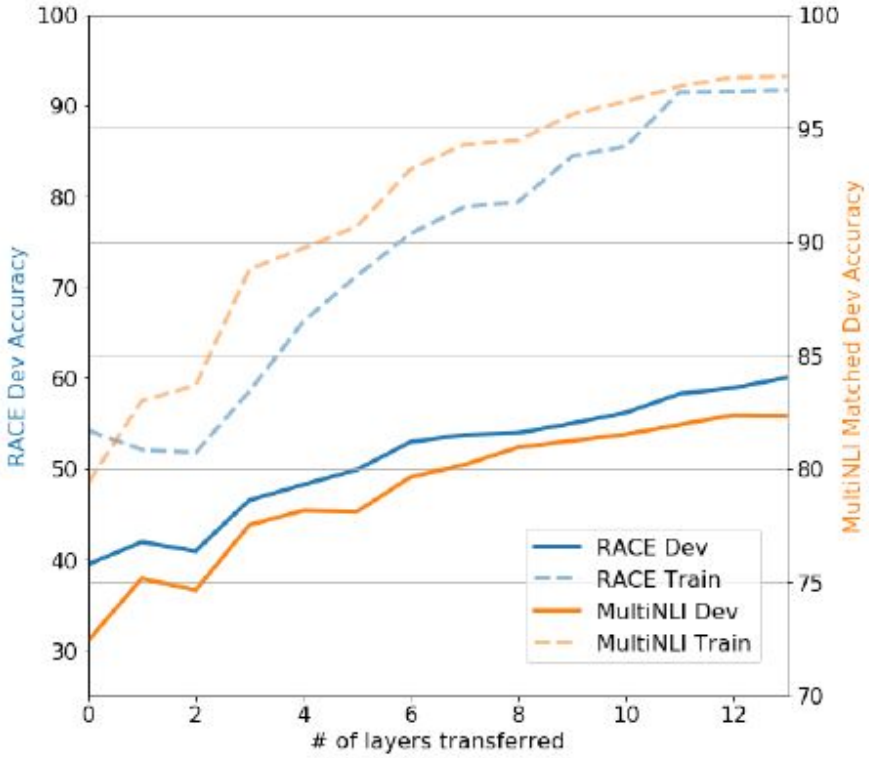
Performance

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

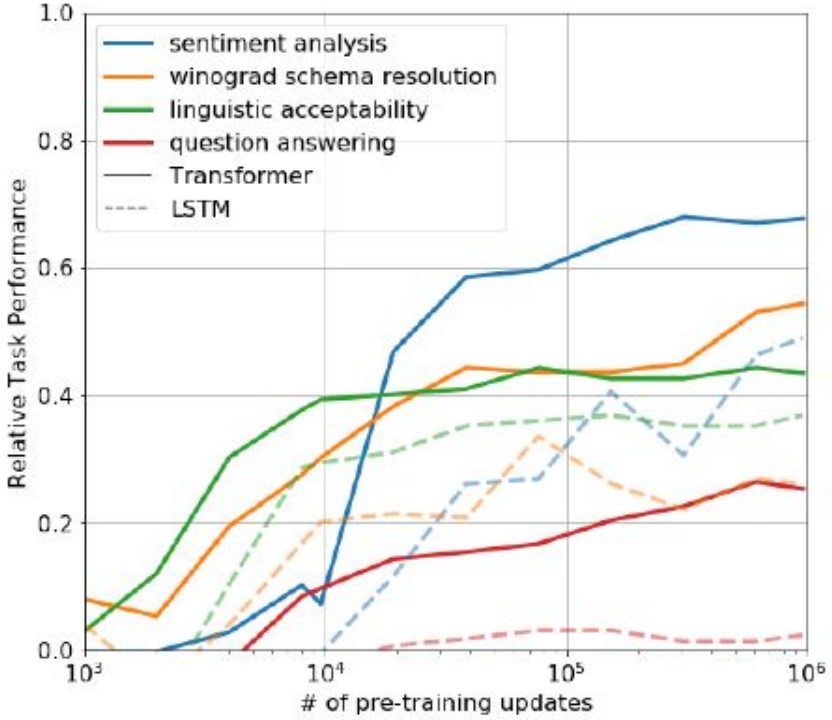
발표 당시 12개 dataset 중에 9개에서 sota 기록

- 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


RACE Dev Accuracy
MultiNLI Matched Dev Accuracy
# of layers transferred
RACE Dev
RACE Train
MultiNLI Dev
MultiNLI Train
100
90
80
70
60
50
40
30
95
85
75
0
2
4
6
8
10
12

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

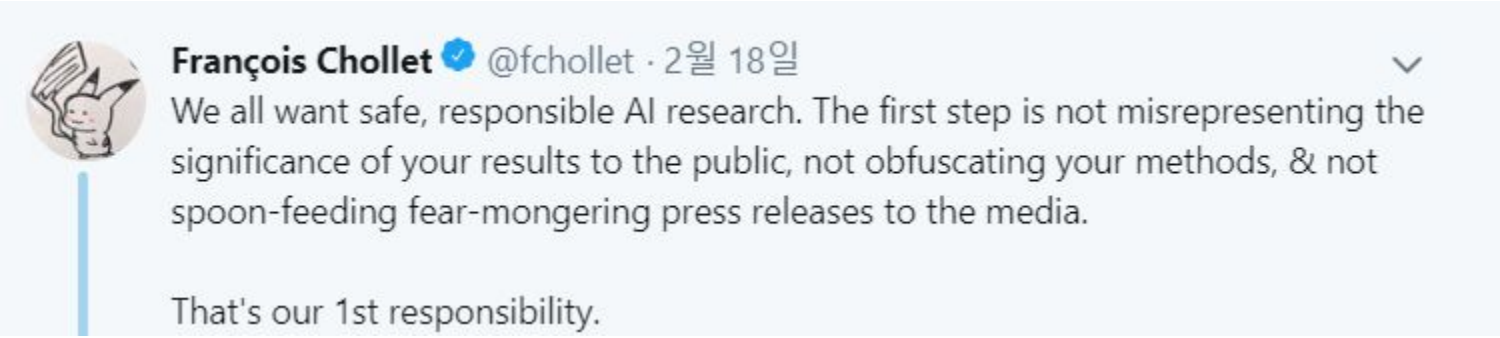
Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

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



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