GPT-1

Improving Language Understanding by Generative Pre-Training
(Technical Report, OpenAI. 2018)

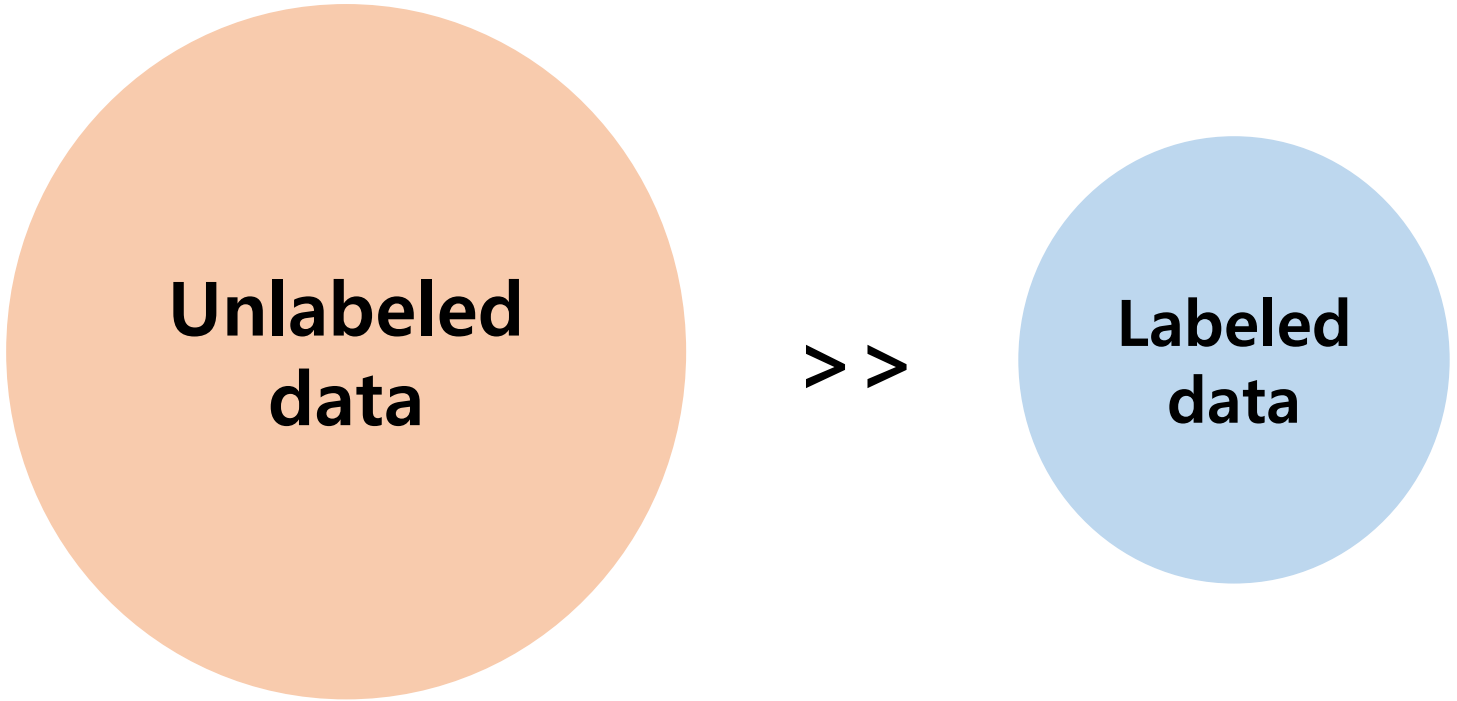
인공지능 연구실 인공지능 세미나
석사 1기 설지우

Contents

1

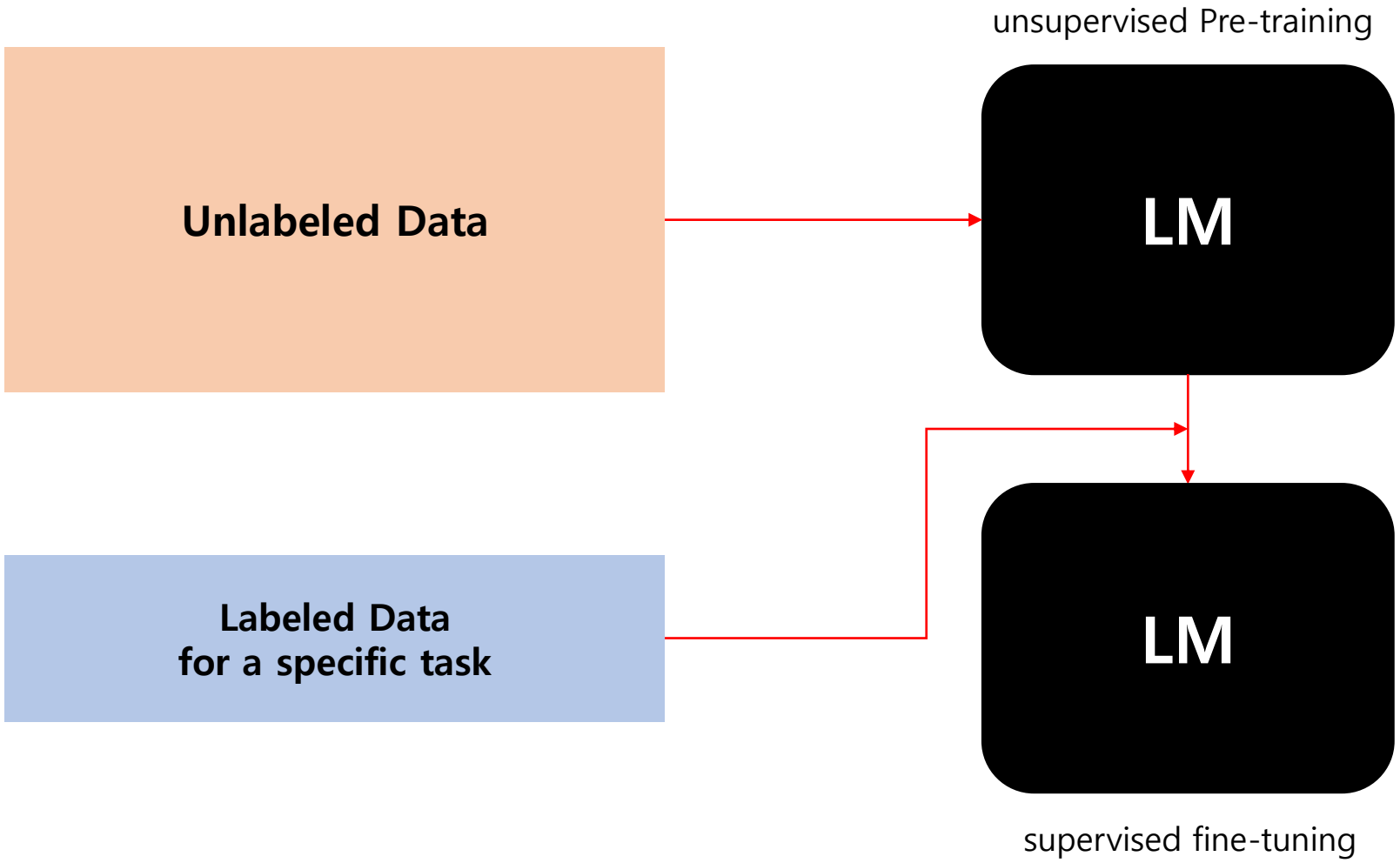
Introduction

1 Introduction



대부분의 deep learning method에서는 상당한 양의 labeled data가 필요했음
→ 하지만, specific task에 필요한 labeled data는 매우 드물었음
→ 그러면 labeled data보다 풍부한 unlabeled data를 잘 활용해보면 supervised task에서도 좋은 성능을 낼 수 있지 않을까?

[Motivation]


unsupervised Pre-training
Unlabeled Data
LM
Labeled Data
for a specific task
LM
supervised fine-tuning

Leveraging more than word-level information from unlabeled data is challenging

- It is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer
  - 단순히 unlabeled data로만 과연 어떠한 목적함수가 효과적인지는 모른다!

- There is no consensus on the most effective way to transfer these learned representations to the target task
  - 각각의 Target task로 transfer learning을 하는데 어떤게 가장 효과적인 방법일지 모른다!

앞의 2가지 불확실성을 보완하는 새로운 semi-supervised approach를 제시한다!

- **two-stage training procedure**를 제안함
  1) Neural Network model의 초기 parameter들을 학습하기 위해 unlabeled data에 대한 language modeling objective를 사용한다
  2) Target task에 상응하는 supervised objective를 사용하여 parameters를 조정한다

2

Model Architecture

2 Model Architecture

[two-stage training procedure]

Step 1

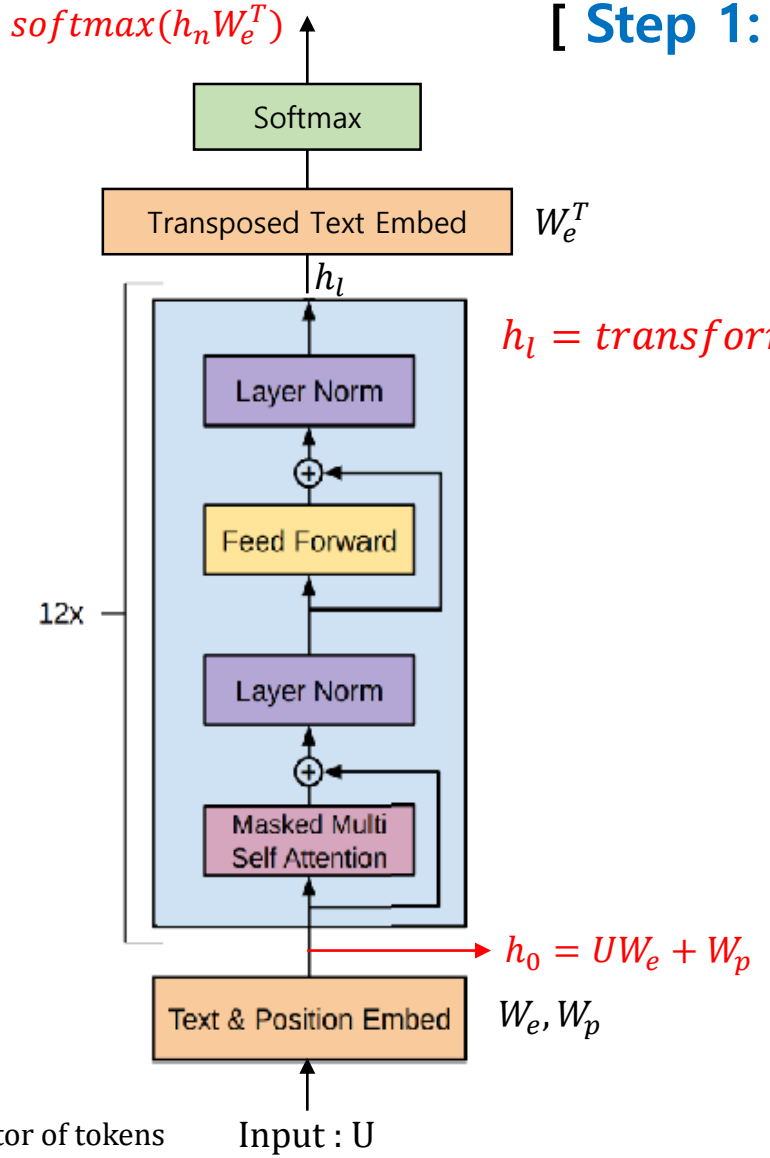
- Unsupervised pre-training
- Unlabeled large data로 학습

→

Step 2

- Supervised fine-tuning
- Specific task에 대한 labeled data로 학습

2 Model Architecture



[Step 1: Unsupervised Pre-Training]

$h_l = transformer\_block(h_{l-1}) \quad \forall l \in [1, n]$

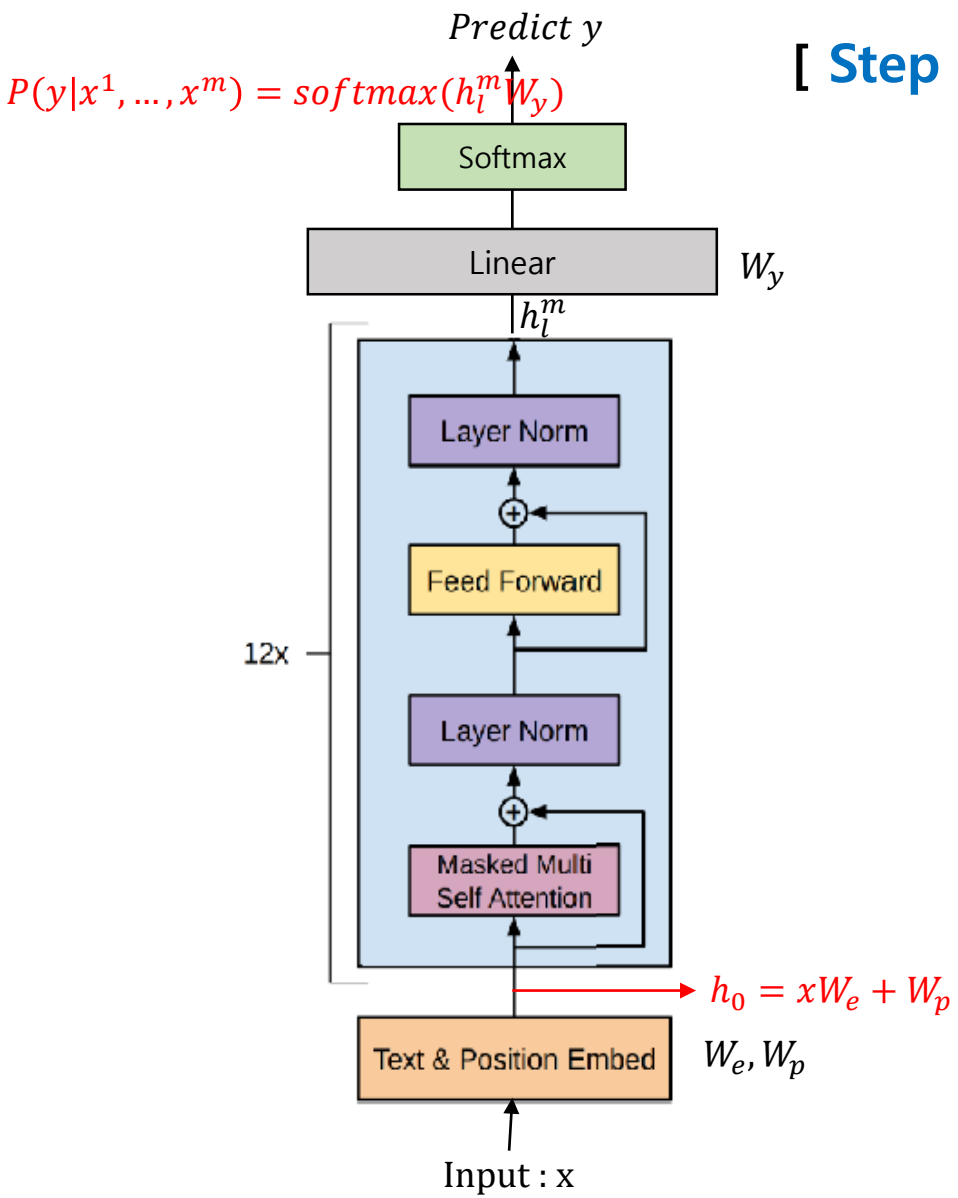
Objective Function

- Unsupervised corpus of token $U = (u_1, u_2, \dots, u_{i-1};\ \Theta)$
- $L_1(U) = \sum_i logP(u_i | u_{i-k}, \dots, u_{i-1}; \Theta)$

→ unsupervised corpus에 대해 window size k개 token이 주어졌을 때, 다음 token을 예측
→ 다음 token이 등장할 likelihood를 최대화하도록 학습함

Q) 왜 window size k개로 지정을 했을까?
A) 검색해보면 기존 decoder랑 똑같음
다른 사람의 GPT PyTorch code도 봤는데 차이가 없음
그래서 저자가 왜 paper에 저렇게 써놓았는지 이유를 모르겠음

2 Model Architecture



[Step 2: Supervised Fine-Tuning]

Ex) a specific task : Classification

- C : a labeled dataset
  - Input tokens : $(x^1, x^2, \dots, x^m)$
  - Label : y

Objective Function

- Supervised corpus of token C= $(x^1, \dots x^m)$
- $objective: L_2(C) = \sum_{(x,y)} logP(y|x^1,\dots,x^m)$
- $auxiliary\ objective: L_1(C) = \sum_i logP(x_i|x_{i-k},\dots,x_{i-1};\Theta)$

⇛ $L_3(C) = L_2(C) + \lambda * L_1(C)$

[Step 2: Supervised Fine-Tuning]

Objective Function

- Supervised corpus of token C= $(x^1, \dots x^m)$
- $objective: L_2(C) = \sum_{(x,y)} logP(y|x^1, \dots, x^m)$
- $auxiliary\ objective: L_1(C) = \sum_i logP(x_i|x_{i-k}, \dots, x_{i-1}; \Theta)$

⇛ $L_3(C) = L_2(C) + \lambda * L_1(C)$

왜 $L_1(C)$를 포함했을까?

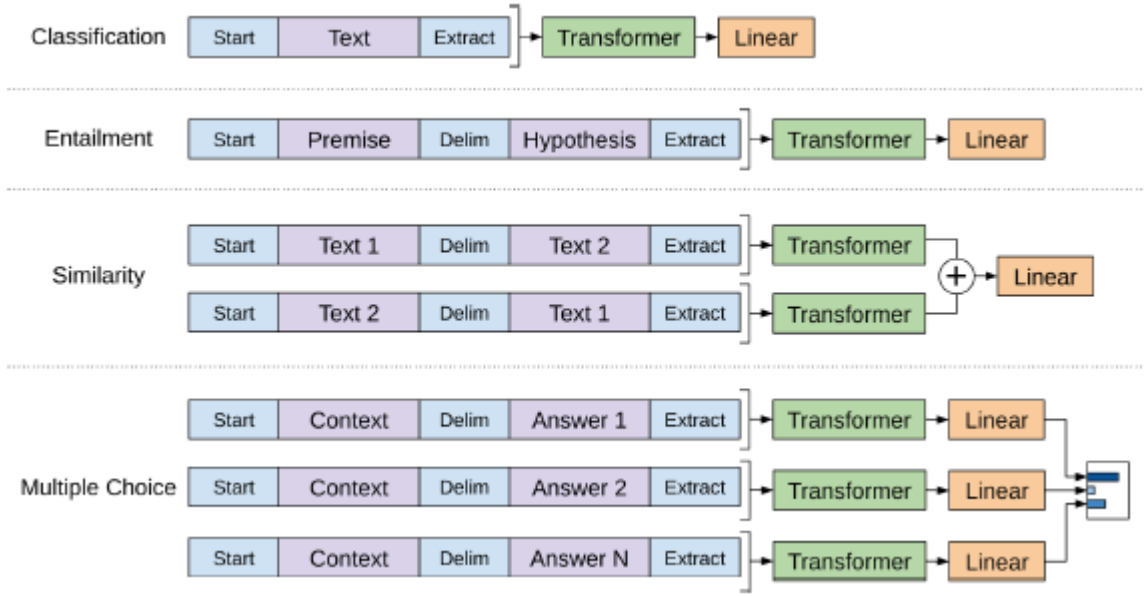
supervised learning에 해당 하는 objective($L_2(C)$) 뿐만 아니라
unsupervised learning에 해당하는 objective($L_1(C)$)도
함께 update되게 되면 2가지 장점을 저자들은 발견했다고 함

- Improving generalization of the supervised model
- Accelerating convergence

[Task-specific input transformations]

- QA, Textual entailment 같은 특정 tasks에 대해서는 추가적으로 input transformations을 수행함
- 이전 연구에는 상당한 양의 task-specific에 대해 새로운 아키텍처를 추가하고 이러한 추가적인 아키텍처에 대해 transfer learning을 사용하지 않음
- 하지만, GPT에서는 structured input을 pre-trained model이 처리할 수 있는 ordered sequence로 바꿔주는 traversal-style approach를 사용!

→ 이 덕분에 방대한 tasks에 대해 아키텍처를 수정하는 걸 하지 않아도 된다!



▲ Figure 1

Task	Input transformation
Textual Entailment	[<s>; premise; $; hypothesis; <e>]
Similarity	[<s>; text1; $; text2; <e>;] [<s>; text2; $; text1; <e>;] → element-wise summation
Question Answering and Commonsense Reasoning	z : a context document q : question $\{a_k\}$: *a set of possible answers* [<s>; z; q; $; a_k; < e >]

- Start token : <s>
- Delim token : $
- Extract token : <e>

3

Experiments

3 Setup

[Unsupervised Pre-Training]

- BooksCorpus dataset

# of books	# of sentences	# of words	# of unique words	mean # of words per sentence	median # of words per sentence
11,038	74,004,228	984,846,357	1,316,420	13	11

[Model specifications]

- N = 12
- h = 12
- $d_{model} = 768$
- $d_{ff} = 3072$
- Adam optimizer (max l_rate = $2.5e^{-04}$)

…

- 나머지 구체적인 사항은 Sec 4.1에서 확인

[Supervised Fine-Tuning]

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Supervised Task Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

→ 4가지 task에 대해 실험을 진행함

[Natural Language Inference]

- 두 문장 간 관계를 맞추는 Task (recognizing textual entailment라고도 함)
- Label은 3가지 종류가 존재 : Contradiction, Neutral, Entailment
- 5가지 Datasets : SNLI, MultiNLI-m, MultiNLI-mm, Question NLI, RTE, SciTail

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

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

- RTE를 제외한 5개의 Dataset에서 SOTA 달성
→ GPT가 다수의 문장들을 잘 추론하고 언어적 모호성을 잘 다룬다
- RTE (2490 examples) 같은 작은 dataset에서는 기존 SOTA 모델보다 accuracy 작음

[QA and commonsense reasoning]

Passage:
In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
"Really? How do you know that?" the gentleman said in surprise.
"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."
The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.
"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.
"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made _.
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because _ .
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that _ .
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by _ .
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made _ .
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

Table 1: Sample reading comprehension problems from our dataset.

- RACE dataset
  - ReAding Comprehension dataset from Examinations
  - 중학교/고등학교 시험에서 문단-질문으로 구성된 dataset
  - CNN이나 SQuAD보다 더 많은 추론(reasoning) 유형 질문을 포함한다
  - 2 subsets : RACE-m (middle school), RACE-h (high school)

- Story Cloze Test dataset
  - 다수의 문장으로 구성된 story에 대한 올바른 엔딩/틀린 엔딩을 추론하는 task

Context	Right Ending	Wrong Ending
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.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.

[QA and commonsense reasoning]

- RACE, Story Cloze dataset에서 모두 SOTA 달성함

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

[Semantic Similarity]

- 주어진 두 문장 간의 유사한 정도를 예측하는 Task
- 3가지 Dataset : MSR Paraphrase Corpus, Quora Question Pairs, STS Benchmark

▼ MSRP Dataset

S. No.	Sentence 1	Sentence 2	Gold annotation	Prediction	Remark
1	Ricky Clemons' brief, troubled Missouri basketball career is over.	Missouri kicked Ricky Clemons off its team, ending his troubled career there.	paraphrase	paraphrase	correct
2	But 13 people have been killed since 1900 and hundreds injured.	Runners are often injured by bulls and 13 have been killed since 1900.	non-paraphrase	non-paraphrase	correct
3	I would rather be talking about positive numbers than negative.	But I would rather be talking about high standards rather than low standards.	paraphrase	paraphrase	correct
4	The tech-heavy Nasdaq composite index shot up 5.7 percent for the week.	The Nasdaq composite index advanced 20.59, or 1.3 percent, to 1,616.50, after gaining 5.7 percent last week.	non-paraphrase	paraphrase	incorrect
5	The respected medical journal Lancet has called for a complete ban on tobacco in the United Kingdom.	A leading U.K. medical journal called Friday for a complete ban on tobacco prompting outrage from smokers groups.	non-paraphrase	paraphrase	incorrect
6	Mrs. Clinton said she was incredulous that he would endanger their marriage and family.	She hadn't believed he would jeopardize their marriage and family.	paraphrase	non-paraphrase	incorrect

▼ STS-B Dataset

Score	English	Spanish
5/4	*The two sentences are completely equivalent, as they mean the same thing.*	
	The bird is bathing in the sink. Birdie is washing itself in the water basin.	El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*	
	In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.	
3	*The two sentences are roughly equivalent, but some important information differs/missing.*	
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.	John dijo que él es considerado como testigo, y no como sospechoso. "Él ya no es un sospechoso," John dijo.
2	*The two sentences are not equivalent, but share some details.*	
	They flew out of the nest in groups. They flew into the nest together.	Ellos volaron del nido en grupos. Volaron hacia el nido juntos.
1	*The two sentences are not equivalent, but are on the same topic.*	
	The woman is playing the violin. The young lady enjoys listening to the guitar.	La mujer está tocando el violín. La joven disfruta escuchar la guitarra.
0	*The two sentences are completely dissimilar.*	
	John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.	Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.

[Classification]

▼ CoLA Dataset

Label	Sentence
*	The more books I ask to whom he will give, the more he reads.
✓	I said that my father, he was tight as a hoot-owl.
✓	The jeweller inscribed the ring with the name.
*	many evidence was provided.
✓	They can sing.
✓	The men would have been all working.
*	Who do you think that will question Seamus first?
*	Usually, any lion is majestic.
✓	The gardener planted roses in the garden.
✓	I wrote Blair a letter, but I tore it up before I sent it.

- CoLA dataset
  - The Corpus of Linguistic Acceptability
  - 문법적으로 맞았는지 틀렸는지를 분류하는 task
- SST-2 dataset
  - The Standford Sentiment Treebank
  - 문장의 sentiment 분류하는 task (일반적인 binary classification task)
  - Label : Positive, Negative

[Semantic Similarity & Classification]

Table 4: Semantic similarity and classification results, comparing our model with current state-of-the-art methods. All task evaluations in this table were done using the GLUE benchmark. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

- **Sentiment Similarity**
  - 3가지 중 2가지 dataset에서 SOTA를 달성함
- **Classification**
  - CoLA는 5.4%로 큰 격차를 보임
  - SST-2에서는 SOTA보다는 1.9% 낮음

[실험결과 요약]

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

→ GPT는 4 specific task에서 12개 datasets 중 9개에서 SOTA를 달성함

→ 작은 dataset (ex. STS-B 5.7k examples)부터 큰 dataset(ex. SNLI 550k examples)까지 다른 크기의 dataset에서도 GPT는 잘 동작하는 걸 확인할 수 있었음

4

Analysis

4 Analysis

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

[Impact of number of layers transferred]

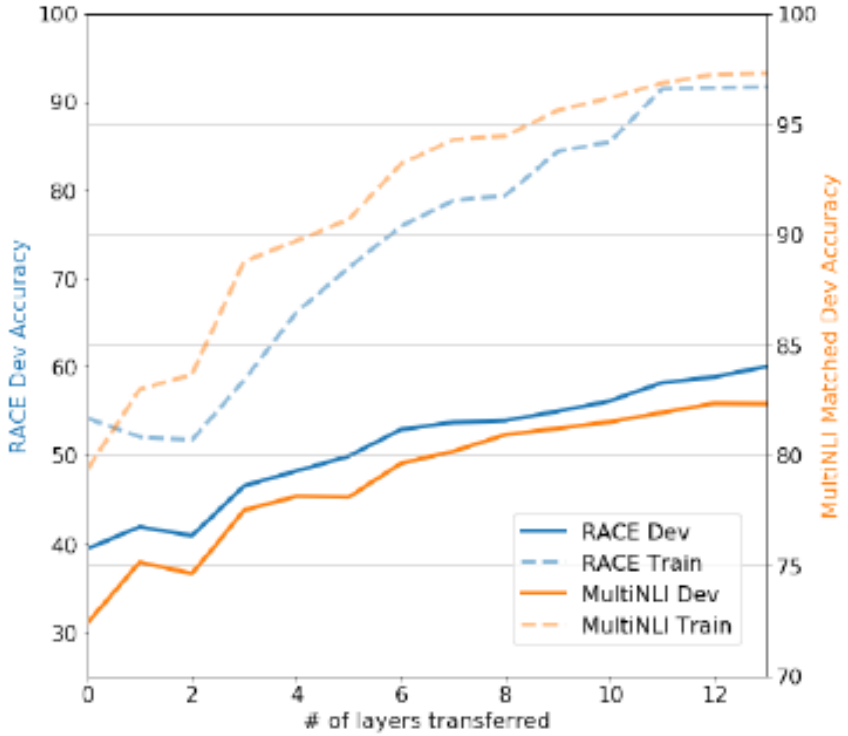


Figure2 (left) : Effect of transferring increasing number of layers from the pre-trained language model on RACE and MultiNLI

- x축 : layer 개수, y축 : accuracy
- Layer의 개수가 증가함에 따라 accuracy가 향상되는 걸 확인할 수 있음
- Layer 12 이후부터는 수렴하는 양상을 보임

4 Analysis

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

[Zero-shot Behaviors]

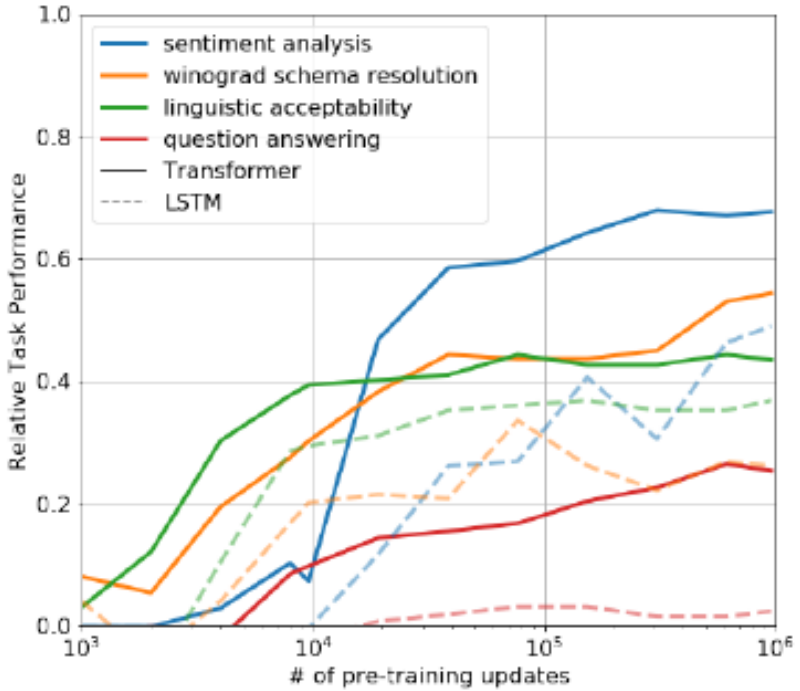


Figure 2 (right) : Plot showing the evolution of zero-shot performance on different tasks as a function of LM pre-training updates

- x축 : # of pre-training updates, y축 : 각 Task에 대한 Performance
- Fine-tuning을 하지 않고 pre-training만한 상태에서 비교 하였을때, GPT가 LSTM보다 성능이 뛰어나다는 걸 보여주기 위해서 이러한 추가실험을 한 것 같음!

- CoLA (linguistic acceptability)
- SST-2 (sentiment analysis)
- RACE (question answering)
- DPRD (winograd schemas)

4 Analysis

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

[3가지 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

- w/o : without, w/ : with
- aux LM : auxiliary LM objective

1) Examine performance of GPT without auxiliary LM objective
   → NLI와 QQP에서 auxiliary LM objective가 효과적이란걸 관찰할 수 있음
   → dataset이 크면 클수록 auxiliary objective가 더 장점이 된다!

2) Analyze the effect of the Transformer by comparing it with a single layer 2048 unit LSTM using the same framework
   → 평균적으로 LSTM이 5.6 score가 Transformer보다 뒤처짐
   → 그래도 MRPC에서는 LSTM이 Transformer보다 더 성능이 뛰어나다

3) Compare with our transformer architecture directly trained on supervised target tasks, without pre-training
   → pre-training이 부족하면 모든 tasks에서 성능이 저하되어 full model에 비해 14.8% 감소한다.

5 Summary

- GPT : Generative Pre-Training Transformer
- Transformer의 decoder만을 사용함
- Generative Pre-Training을 통해 task-agnostic model을 제시

END

[Zero-shot Behaviors]

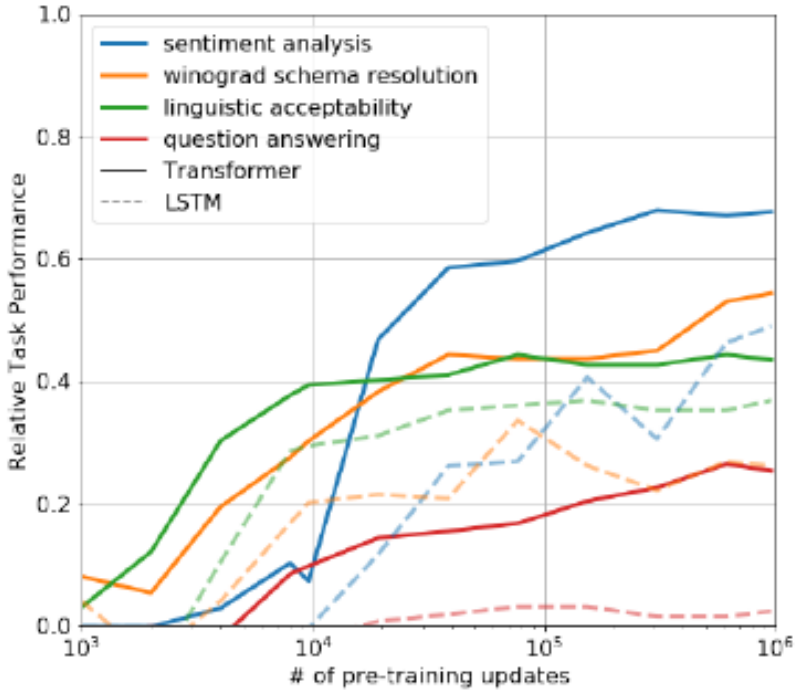


Figure 2 (right) : Plot showing the evolution of zero-shot performance on different tasks as a function of LM pre-training updates

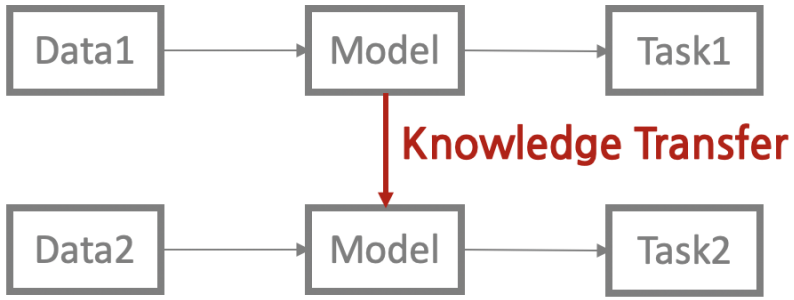
- CoLA (linguistic acceptability) : average token log probability로 scoring됨
- SST-2 (sentiment analysis) : 각 예시에 very token을 추가하고 label을 positive/negative로 한정됨
- RACE (question answering) : generative model이 가장 높은 average token log-probability를 부과한 answer를 고름
- DPRD (winograd schemas) : definite pronoun(ex. he, she 등)을 2개의 referrents로 대체함. 그 후 나머지 시퀀스에 더 높은 average token log-probability를 부과한 resolution을 예측함

Transfer learning

[기존의 문제점]
: 딥러닝 모델을 제대로 훈련시키려면 많은 수의 데이터가 필요함
-> 하지만 충분히 큰 데이터셋을 얻는 것은 어려운이 있음

[해결책 : 전이학습]
전이학습이란 이미지넷과 같이 아주 큰 데이터셋에 훈련된 모델의 가중치를 가지고 와서 해결하고자 하는 task에 맞게 재조정해서 사용하는 것을 말한다
즉, 다른 분야(task)에서 훈련된 신경망(pre-trained model)을 새로운 분야 혹은 비슷한 분야(task)의 신경망 학습에 재사용(reuse)하는것

[장점]
새로 모델을 구성하여 학습시키는 것보다 시간/비용적인 측면에서 효율적이다
Pre-trained 모델은 적용하려는 task의 데이터가 학습할 때의 데이터와 같은 분포를 가진다고 가정할 시, 효율적이다



Task1 : upstream task – 다음 단어 맞추기
Task2 : downstream task - classification

Downstream task를 학습하는 방식

- fine-tuning : downstream task data 전체를 사용. 해당 task에 맞게 모델 전체를 update
- Prompt tuning : downstream task data 전체를 사용. 해당 task에 맞게 모델 일부를 update
- In-context learning : downstream task data 일부만 사용. Model을 update하지 않음

In-context learning에서의 3가지 방식

- zero-shot learning : downstream task data를 전혀 사용하지 않는다. 모델이 바로 downstream task를 수행한다
- One-shot learning : downstream task data를 하나만 사용한다. 모델은 1건의 data가 어떻게 수행되는지 참고한 뒤 downstream task를 수행한다
- Few-shot learning : downstream task data를 몇 건(few)만 사용한다. 모델은 몇 건의 data가 어떻게 수행되는지 참고한 뒤 downstream task를 수행한다

[Generative Model]

- P(X|θ)를 P(θ)와 P(θ|X)를 활용해서 간접적으로 도출

$$P(X|\theta) = \frac{P(\theta|X)P(X)}{P(\theta)}$$

- 데이터가 생성되었을 가능성이 높은 분포의 parameter θ를 찾는 과정

[Discriminative Model]

- 데이터 X가 주어졌을 때 label Y가 나타날 조건부 확률 p(Y|X)를 반환하는 모델
- X의 label을 잘 구분하는 decision boundary를 학습하는 것이 목표
- Generative model에 비해 가정이 단순하고 학습데이터 양이 충분하다면 좋은 성능을 낸다

모르는 용어

[word-level information]
- Ex) word-embeddings, which are trained on unlabeled corpora
- These approaches, however, mainly transfer word-level information, whereas we aim to capture higher-level semantics

[traversal-style approach]
- Which process structured text input as a single contiguous sequence of tokens
- Where we convert structured inputs into an ordered sequence that our pre-trained model can process

[Task-agnostic model]
- General purpose model
- Task에 구애받지 않는

[zero-shot behaviors]
- It acquires useful linguistic knowledge for downstream tasks
- Test 때 Learner는 훈련 중에 관찰되지 않은 class의 샘플을 관찰하며 이들이 속한 범주를 예측해야 함

Winograd schema

Winograd schema : 매우 모호한 대명사(pronoun)을 가진 하나 또는 두 단어가 무엇을 의미하는지 선택하는 문제 (commonsense가 필요함)

1. *The city councilmen refused the demonstrators a permit because* ***they*** *feared violence.* Who feared violence?
   A. **The city councilmen** B. The demonstrators

2. *The city councilmen refused the demonstrators a permit because* ***they*** *advocated violence.* Who advocated violence?
   A. The city councilmen B. **The demonstrators**

3. *The trophy doesn't fit in the brown suitcase because* ***it*** *is too big.* What is too big?
   A. **The trophy** B. The suitcase

4. *The trophy doesn't fit in the brown suitcase because* ***it*** *is too small.* What is too small?
   A. The trophy B. **The suitcase**