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

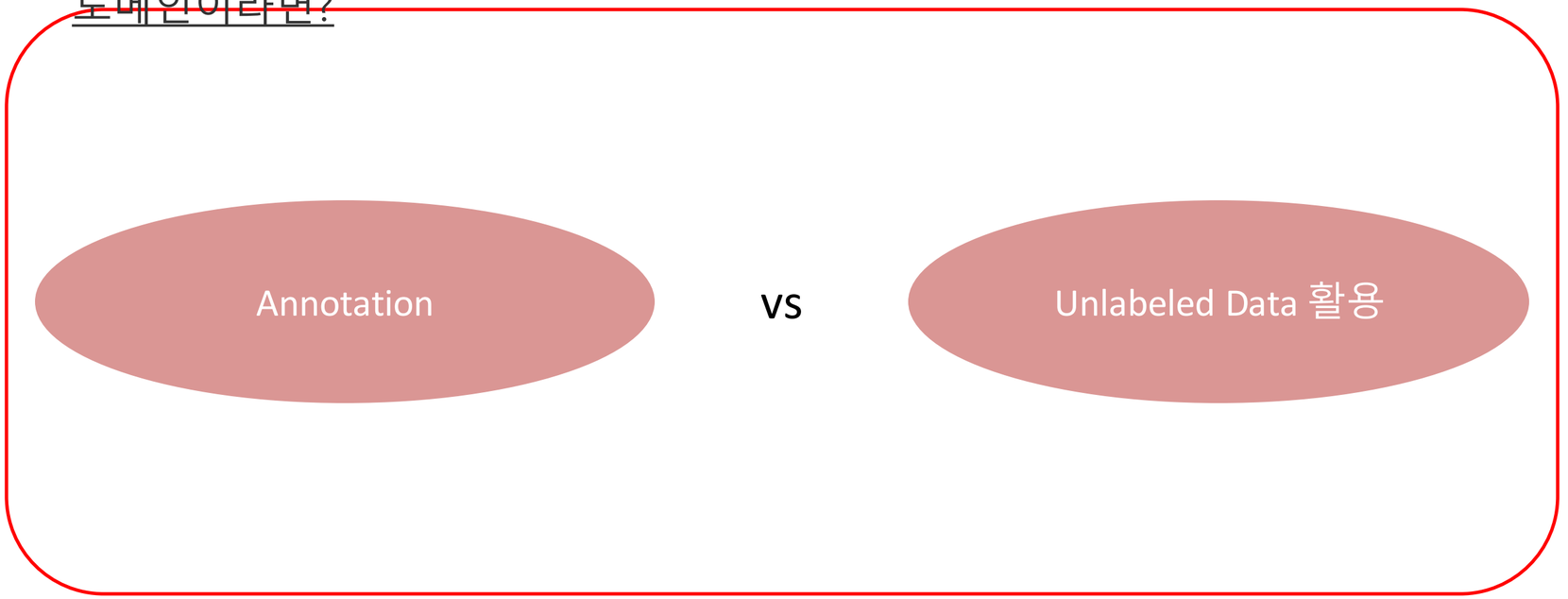
Hwang Hyeon Tae

01

Introduction

Unsupervised Learning

- 대부분의 Deep Learning 방식
  - 다량의 labeled Data가 필수
  - 하지만 리소스가 부족한 도메인이라면?



Unsupervised Learning

- Annotation
  - 시간과 비용이 너무 많이 듦

- Unlabeled Data 활용
  - Unlabed Data의 언어 정보를 활용할 수 있는 언어 모델을 만들 수 있다면 Annotation의 단점 해결 가능
  - Supervision이 가능한 경우에도 Unsupervision 방식으로 좋은 표현을 학습할 수 있다면 성능 향상이 가능
    - ➢ pre-trained word embedding을 사용해 다양한 NLP task의 성능 개선이 그 증거

Unlabeled Data

- Unlabeled text에서 단어 수준 이상의 정보 활용이 어렵다.

1. 어떤 optimization objective가 transfer에 유용한 텍스트 표현을 학습하는데 효과적인지 불분명
   - 단어 수준
     - Word2Vec이나 Glove 같은 모델은 단어의 분산 표현을 학습하기 위해 예측 기반의 optimization objective를 사용
     - 따라서 주변 단어를 기반으로 특정 단어를 예측하거나, 단어 쌍 간의 관계 학습 가능
   - 단어 이상의 수준(문장 이상)
     - 더 큰 범위의 경우 주변 단어를 예측하는 것만으로는 충분 X
     - 따라서 어떤 optimization objective를 사용해서 모델을 학습해야 할 지 파악이 어려움

" 어떤 표현을 학습해야 하는가"

Unlabeled Data

- Unlabeled text에서 단어 수준 이상의 정보 활용이 어렵다.

2. 학습된 표현을 target task에 가장 효과적으로 transfer하는 방법에 대한 합의가 이루어지지 않았다.
   - 1번의 문제를 해결했다고 가정할 때
     - 그렇게 학습된 표현은 또 어떻게 Downstream에 적용해야 하는지 모르겠다.

" 학습된 표현을 어떻게 활용해야 하는가”

Unlabeled Data

이러한 불확실성 때문에 언어 처리를 위한
효과적인 Semi-supervised learning approach 개발에 어려움

GPT-1

- 2단계 training procedure을 활용한 semi-supervised approach 소개

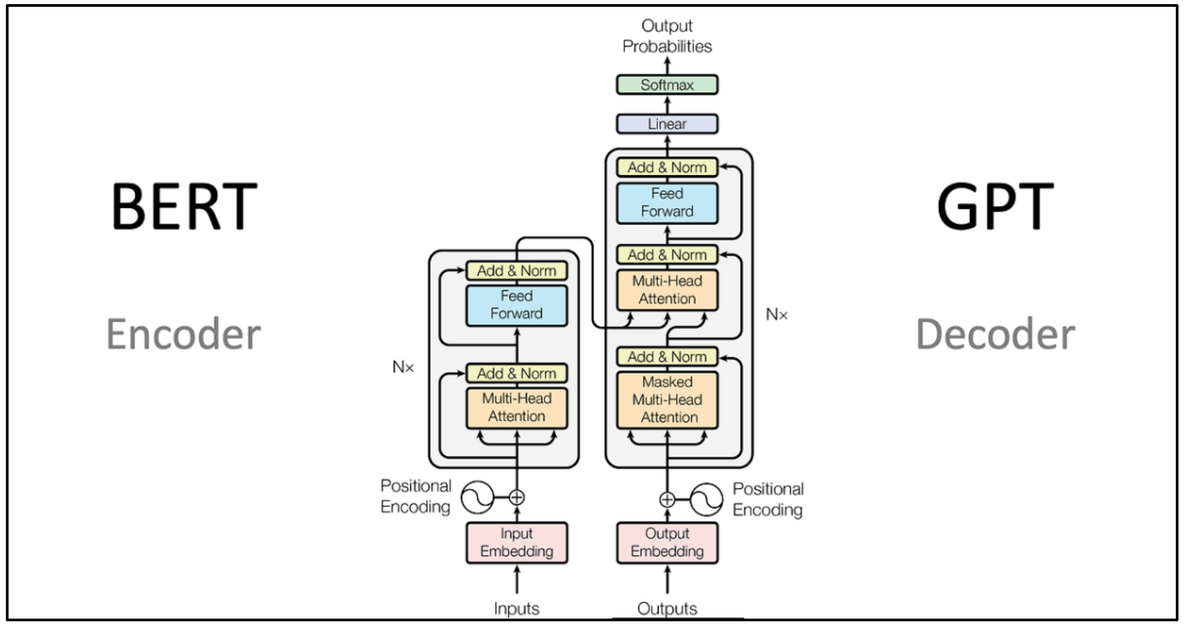
1. Unsupervised pre-training
   - Unlabeled Data에 language modeling objective를 사용해 신경망 모델의 초기 파라미터 학습

2. Supervised fine-tuning
   - 해당 Supervised objective를 사용해 파라미터를 Target Task에 맞게 tunin한다.

Target Task가 Unlabeled Data와 같은 도메인이 아니어도 된다.

GPT-1

- Model Architecture : Transformer
  - RNN, LSTM과 같은 Recurrent Network와 비교해 long-term dependency 문제에서 더 Robust한 transfer 성능 제공



02

Framework

GPT-1

- 2단계 training procedure을 활용한 semi-supervised approach 소개

1. Unsupervised pre-training
   - Unlabeled Data에 language modeling objective를 사용해 신경망 모델의 초기 파라미터 학습

2. Supervised fine-tuning
   - 해당 Supervised objective를 사용해 파라미터를 Target Task에 맞게 tunin한다.

Target Task가 Unlabeled Data와 같은 도메인이 아니어도 된다.

Unsupervised pre-training

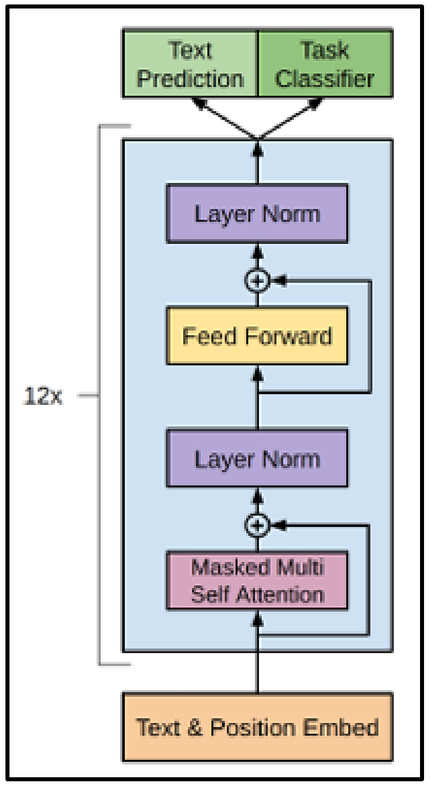
- Loss function
  - Maximum Likelihood Estimation
    - $\mathcal{L}_1(v) = \sum_i logP(u_i|u_{i-k}, \dots, u_{i-1}; \theta)$
- Optimization Algorithm
  - SGD(Stochastic Gradient Descent) 기반의 파라미터 업데이트

Note

K : Context Window Size
u_i : 현재 예측하고자 하는 단어
θ : 모델 파라미터

Unsupervised pre-training

- In experiment,
  - Multi-layer Transformer decoder를 Language Model로 사용
  - 과정
    1. Input context token에 대해
       - ➢ $h_0 = UW_e + W_p$
    2. Multi-headed Self-Attention을 적용한 다음,
       Position-wise Feed Forward layers를 적용해
       - ➢ $h_l = transformer\ block_{(h_{l-1})\forall_i} \in [1, n]$
    3. Target Token에 대한 Output distribution을 생성한다.
       - ➢ $P(u) = softmax(h_n W_e^T)$

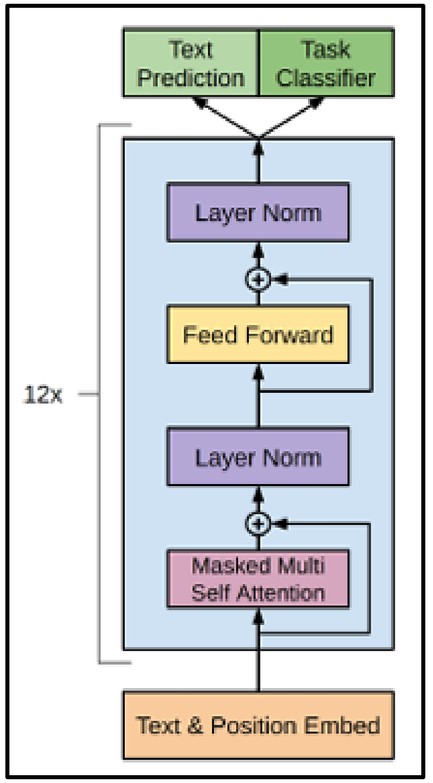


Note

U : Context Vector
n : 레이어 수
W_e : Token Embedding Matrix
W_p : Position Embedding Matrix

Supervised fine-tuning

- Parameter를 Target Task에 맞게 tuning
  - Input
    - h_l^m : Pre-trained model 마지막 transformer block의 activation
    - W_y : Parameter
  - Output Distribution
    - $P(y|x^1, \dots, x^m) = softmax(h_l^m W_y)$



Supervised fine-tuning

- Parameter를 Target Task에 맞게 tuning
  - Loss function

$$\mathcal{L}_2(v) = \sum_{(x,y)} logP(y|x^1, \dots, x^m)$$

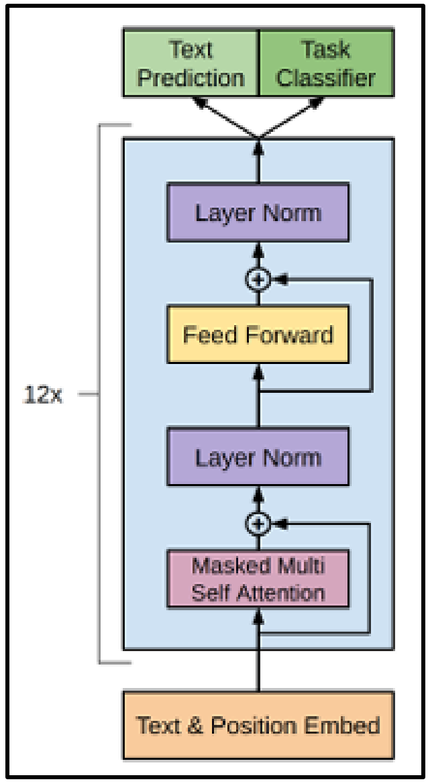
- 이 과정에서,

  language model을 fine-tuning의 auxiliary objective로 설정할 경우

  1. Supervised Model의 generlization을 개선하고
  2. convergence를 가속화한다.

  - 구체적인 Loss function

$$\mathcal{L}_3(C) = \mathcal{L}_2(C) + \lambda * \mathcal{L}_1(C)$$



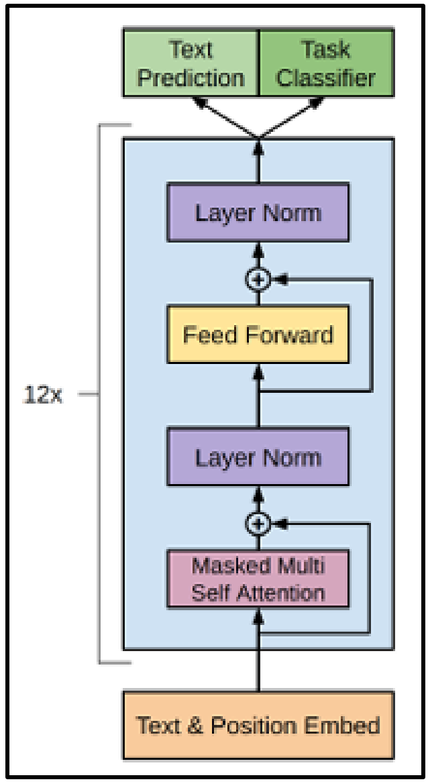
03

Experiment

Setup : Unsupervised pre-training

- Model specifications
  - Transformer Block : 12-layer decoder-only transformer
    - masked self-attention heads 포함

      (768 dimentional states and 12 attention head)
    - position-wise feed-forward networks
      - 3072 demensional inner state
  - Optimizer : Adam
  - Scheduler : Cosine annealing learning rate scheduler
  - Activation function : GELU



Unsupervised pre-training

- Token level perplexity : 18.4
  - BooksCorpus Dataset 사용
    - 연속된 긴 텍스트 포함

      -> Generative Model이 long-range information 학습

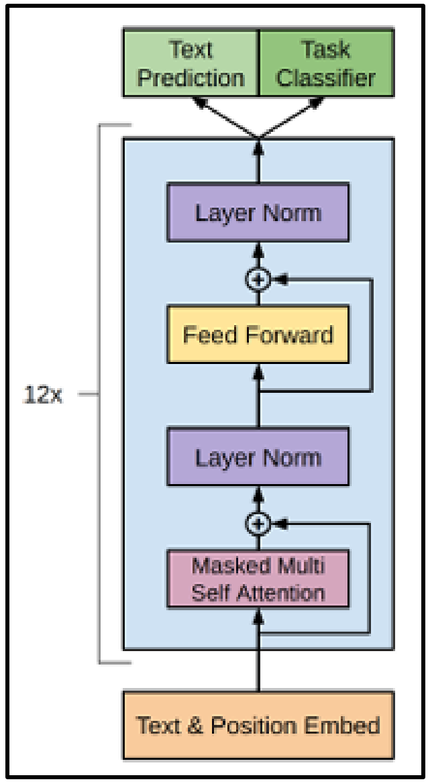
문장들이 연속된 하나의 시퀀스로 제공

- 유사 접근 방식 ELMo Model에서 사용한 Dataset의 경우는 문장 수준에서 섞이기 때문에 long-range structure 파괴

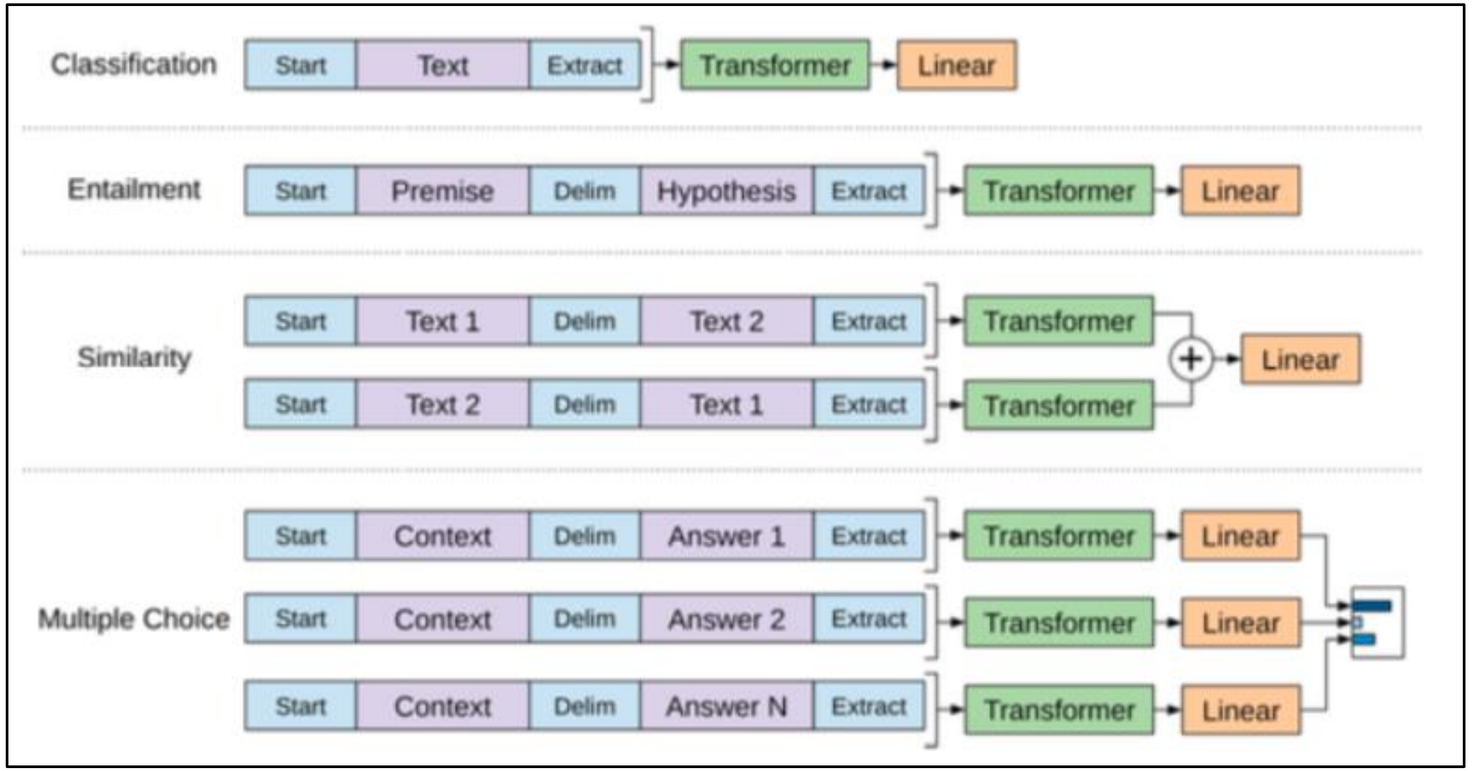
ELMo의 경우 문장 단위로 학습하므로 문장 간의 순서가 중요 x
즉, 여러 문장들이 독립적인 샘플로 구성

Setup : Fine-tuning details

- Model specifications
  - unsupervised pre-training의 hyperparameter setting 재사용
  - Classifier
    - Dropout : 0.1
    - Learning Rate : 6.25e-5
    - batchsize : 32
    - Scheduler : linear learning rate decay schedule

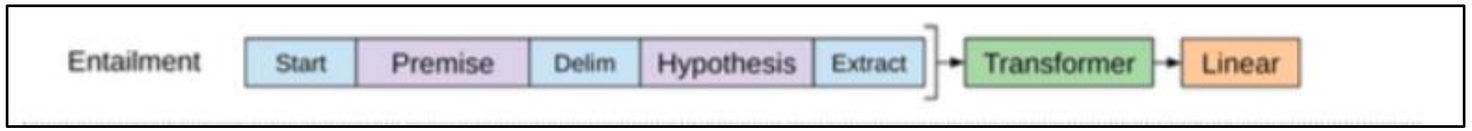


Supervised Fine-tuning



- Task
  1. Natural Language Inference
  2. Question Answering and Commonsense reasoning
  3. Semantic Similarity
  4. Classification

Supervised Fine-tuning : NLI



- NLI(Natural Language Inference)
  - 한 쌍의 문장을 읽고 entailment, contradiction, neutral 중 하나로 문장 간 관계를 판단

Example

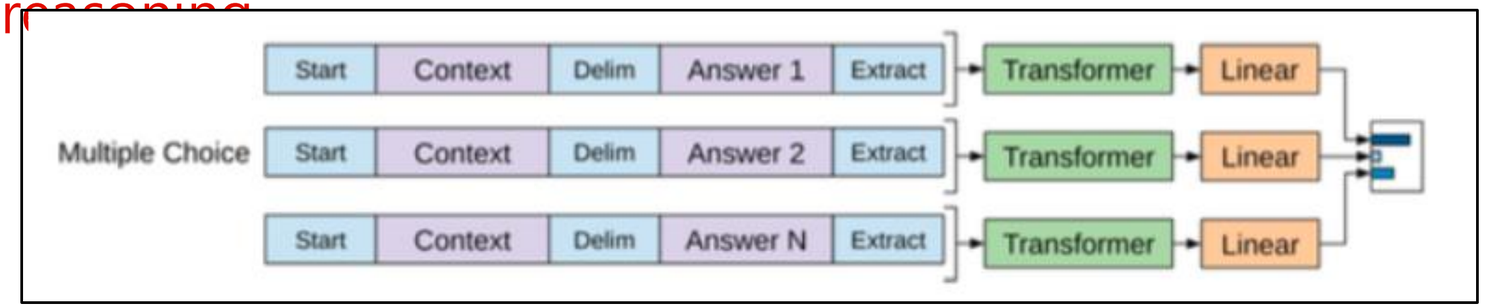
- Premise(전제) :
  "The woman is playing the piano at a concert"
- Hypothesis 1 :
  "A woman is performing music" <- Entailment
- Hypothesis 2 :
  "The woman is painting a picture" <- Contradiction
- Hypothesis 3 :
  "A woman is participating in a cultural event" <- Neutral

Supervised Fine-tuning : NLI

- 평가 데이터 : MNLI(연설, 대중 소설, 정부 보고서), SNLI(이미지 캡션), SciTail(과학 시험), QNLI(위키피디아 기사), RTE(뉴스 기사)
- Result :
  - 5개의 데이터셋 중 4개의 데이터셋에서 크게 능가
  - MNLI에서 최대 1.5%, SNLI에서 0.6%, SciTail에서 5%, QNLI에서 5.8%의 성능 개선
  - "여러 문장을 더 잘 추론하고 언어적인 모호성의 측면을 더 잘 처리한다"

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

Supervised Fine-tuning : Question answering and commonsense reasoning



- Question answering and commonsense reasoning
  - 목표 : 장문의 문맥을 이해하고 처리할 수 있는가?

Example

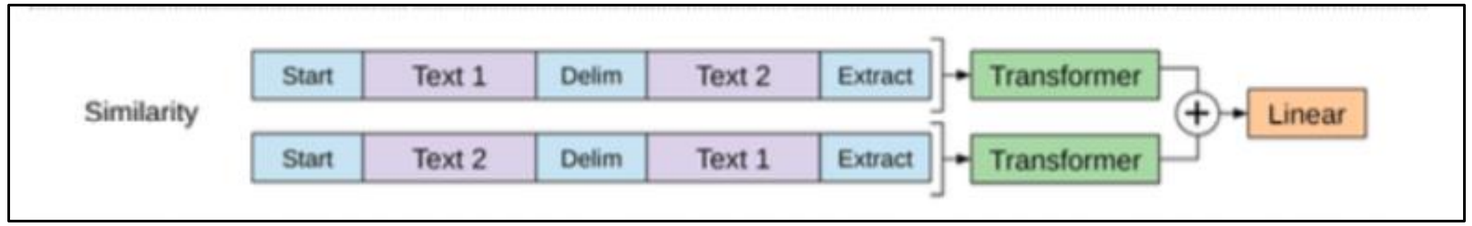
- 문맥이 주어진다.
- Question : "Why did Alice decide to make the scarf herself?”
- Answer :
  - A) She thought it would be fun.
  - B) The scarf in the shop was not beautiful.
  - C) The scarf in the shop was too expensive.
  - D) Her mother asked her to make one.

Supervised Fine-tuning : Question answering and commonsense reasoning

- 평가 데이터 : Story Cloze(스토리의 올바른 결말을 선택하는 데이터셋), RACE(중고등학교 시험의 영어 구절 관련 질문 데이터셋)
- Result :
  - Story Cloze에서는 최대 8.9%, RACE에서는 전체 5.7%의 성능 개선
  - "Long-range context를 더 효과적으로 처리할 수 있다."

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**

Supervised Fine-tuning : Semantic Similarity



- Semantic Similarity
  - 목표 : 두 문장의 의미적으로 동일한가?

Example

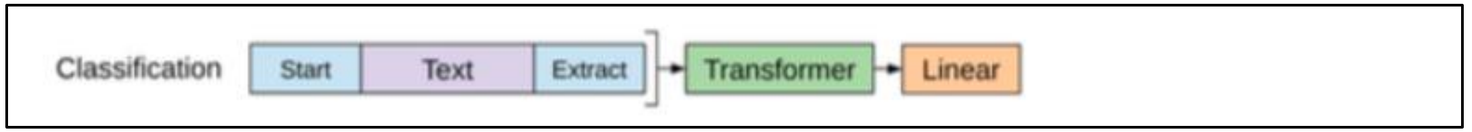
- Text1 : “The cat is sitting on the mat.
- Text2 : “A cat is resting on a rug”

Supervised Fine-tuning : Semantic Similarity

- 평가 데이터 : MRPC(Microsoft Paraphrase corpus), STSB(Semantic Textual Similarity benchmark), QQP(Quora Question Pairs)
- Result :
  - 3가지 Semantic similarity task 중 2가지 task에서 성능 개선
  - "문장 간의 의미를 더 잘 이해할 수 있다."

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**

Supervised Fine-tuning : Classification



- 평가 데이터 : CoLA(문장이 문법적 수용성 평가 데이터), SST2(문장 감성 분석 데이터)
- Result :
  - CoLA에서 45.4점으로 이전 최고 결과인 35.0점보다 크게 상승
  - GLUE benchmark에서도 72.8점으로 이전 최고 결과인 68.9점보다 크게 향상

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**

04

Conclusion

GPT-1

- Generative pre-training과 discriminative fine-tuning을 통해 single task-agnostic model로 robust한 언어 이해를 위한 프레임 워크 도입
  - Experiment를 통해 discriminative task를 성공적으로 해결해 12개의 데이터셋 중 9개의 데이터셋에서 성능 향상

- Unsupervised (pre)training을 사용해 discriminative task의 성능을 향상시키는 것은 오랫동안 머신러닝 연구의 중요한 목표
  - 이번 연구는 실제로 상당한 성능 향상이 가능함을 시사
  - 이를 통해 자연어 이해 및 기타 영역 모두에서 Unsupervised learning에 대한 새로운 연구가 활성화되고 이해도가 더욱 향상되길 바람

Thank you

N. Hwang Hyeon Tae | E. gusxo3975@naver.com |
L..linktr.ee/oneul_