제목: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- 1. Introduction
  - Language model pre-training 효과적으로 Sentence level task(natural language inference, paraphrasing)과 Token-level(NER, SQuAD question answering)를 향상시켰다.
  - Transformer :
    - Transformer Language Model 을 학습
    - 이후 Fine-Tuning 에서 적절히 학습하면 좋은 성과를 낸다
  - Pre-training Language representations to down stream task 에 적용하기 위한 전략 2 개
    - Feature-based
      - ELMo: task-specific architectures 에 추가 적인 Features 로 사용. pre-training 된 representations 들을 포함
    - Fine-tuning
      - Pre-trained parameters 부터 down-stream task 를 simply fine-tuning 하는 것
  - 이전 Pre-training 은 같은 경우 공통 점이 있다.
    - General language representations 를 학습하기 위해서 같은 Objective Function 을 share 하는 것
    - Unidirectional language models 를 사용
      - ELMo 에서는 서로 다른 반향 모델을 학습하면서 합침
  - 이 논문에서는 Unidirectional 한 모델이 전체 pre-training 과 fine-tuning 을 제약한다고 주장
    - 일단 기계번역과 같이 Decoding 에서 단반향으로 선택해서 번역되어야 하지만, Encoding 에서는 적절히 잘 Decoding 에 사용되는 Input 을 잘 만들어내면 된다.. 반향이 중요하지 않다.
    - 예를 들면 OpenAI GPT 에서 left 와 right language architecture 를 사용하지만 모든 Token 은 오직 Previous token 만 attention 한다.
  - 따라서 이 논문은 BERT 를 제안함으로써 Fine-tuning based approach 를 향상시키겠다
    - BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
    - 마지막 Classification Layer 만 추가하고 pre-training 과 fine-tuning 을 추가한다.
    - 처음 부터 Target 을 학습하지 않고, General 한 Corpus 로 사용
    - 이전 Transformer 와 다른 점이 멀까?

- Scare -> 몇 백개 GPU 사용
- Pre-training 을 할 때, MLM 이 학습이 잘된다.
- Large corpus 를 Pre-training 을 함
- Pre-training 시간에 비해서 fine-tuning 이 굉장히 짧다.
    - Epoch 3?

- 새로운 Pre-training objective 를 Proposing 하는데 이것을 Masked Language Model(MLM)이라 함
    - MLM:
        - Token 을 일부를 랜덤하게 Mask 를 씌우고 network output 에서 Mask 의 Original vocabulary 를 맞춤
        - 이 것의 특징은 오직 문맥 정보만을 based 하게 추론해서 맞춤
        - 또한 Left 와 Right 정보를 둘다 썩어서 Deep Bidirectional Transformer 를 Pre-Training 함
    - 추가적으로 Next sentence prediction Task 도입
        - Task: Jointly pre-trains text-pair representations
- Contribution
    - Deep bidirectional pre-training 을 적용하고
    - 독립적으로 학습된 Left-Right , Right-Left LM 의 Shallow concatenation 을 도입
    - 최고의 Performance 적용
    - Eleven LNP task 에 General 하게 모두 적용해서 좋은 결과를 얻음

- 2. Related Work
    - 읽어 보시길
- 3. BERT
    - 3.1 Model Architecture
        - Multi-Layer bidirectional Transformer encoder
        - Configure
            - Number of layers as L
            - Number of self-attention heads as A
            - Number of Hidden size as H
        - BERT_BASE : L=12, H=768, A=12, Total Parameters = 110M
        - BERT_LARGE: L=24, H=1024, A=6, Total Parameters = 340M
        - BERT_BASE same Parameter with OpenAI GPT for comparison

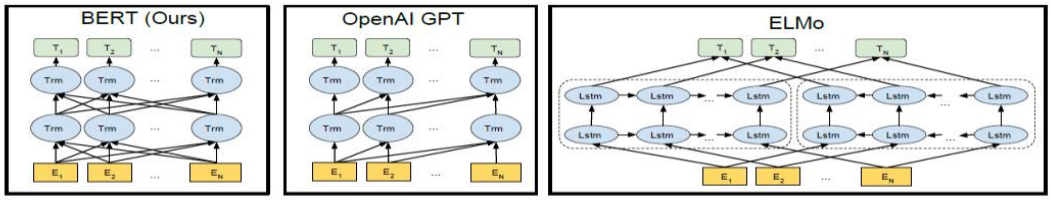


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

- 3.2 Input Representation
  - Pair sentence 와 single 이 다르다.
  - Embedding WordPiece embedding 사용(30,000 token vocabulary)
    - Split word piece -> "##"
  - Learned positional embedding 사용 512 token 사용

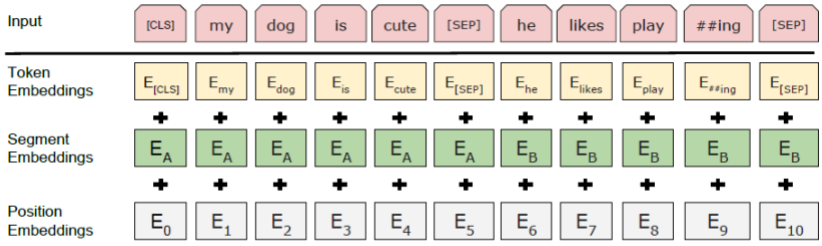


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

- 3.3 Pre-Training task
  - BERT 는 2 개의 Unsupervised prediction task 를 사용
    - T 시점에서 단어 예측에 필요한 Task 를 Pre-Training 하기 위해 사용
    - 3.3.1 Task#1: Masked LM
      - WordPiece token 에 15%를 Random 하게 Mask 를 만들고, 오직 masked word 를 예측한다.
      - 2 개의 downsides
        - 1. Mismatch between pre-training and fine-tuning
          - Pre-training 에서 [MASK] Token 사용. 하지만 Fine-Tuning 에서는 사용하지 않음.
            - 왜냐면 Fine-Tuning 은 다른 Task NLP, NLU …
          - 이를 완화하기 위해서 항상 "masked"데이터를 실제 [MASK] Token 으로 변형하지 않음

- 대신에 training data generator 가 15% 확률로 랜덤하게 Token 을 지정
- Ex) my dog is hairy -> hairy 선택
- 80% of the time [Mask] 교체
  - My dog is hairy -> my dog is [Mask]
- 10% of the time Replace random word
  - My dog is hairy -> my dog is [apple]
- 10% of the time unchanged
  - My dog is hairy -> My dog is hairy
- 하지만 이것은 distributional contextual representation 을 강력하게 유지 하는 효과를 가진다. 왜냐면 특정 비율로 random token 이 발생(Mask, Random word, Un-change)

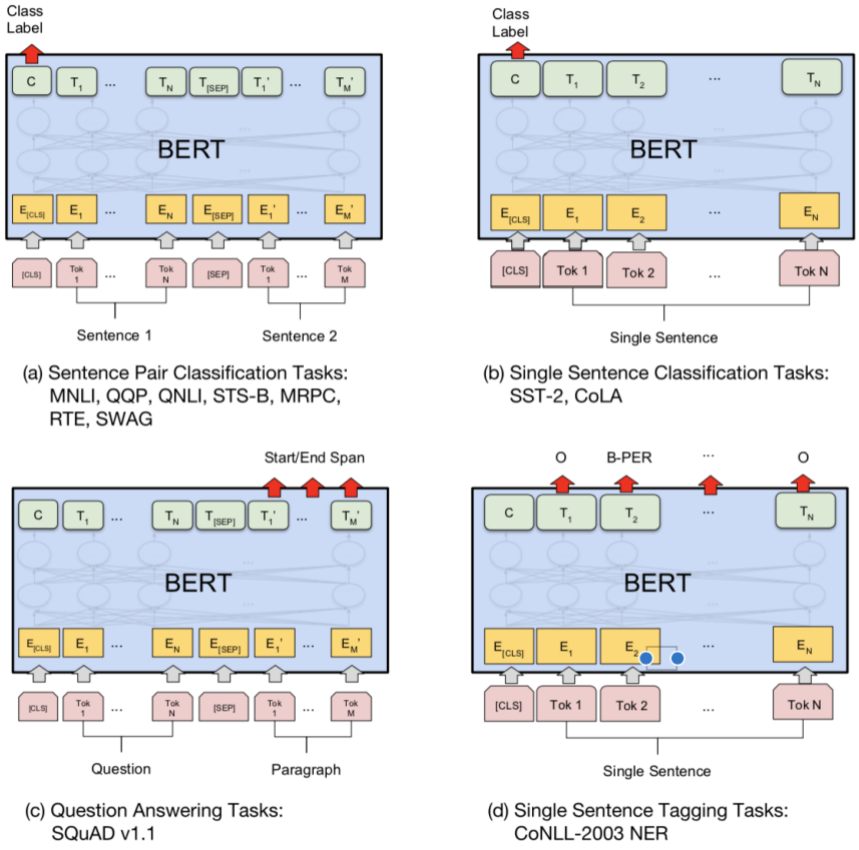
  - MLM 을 사용하면 15% 토큰만 예측하는 Task 를 가지기 때문에 정확성을 위해서 더 많은 Step 에 Training 을 해야한다. 따라서 Left to Right 모델보다 학습 시간이 길지만 경험적으로 훨씬 좋은 Performance 를 가짐

- Task#2: Next Sentence Prediction
  - QA 와 NLI 는 2 문장의 이해 관계를 바로 학습 하기 때문에 어려운 문제이다.
  - 하지만 이런 Task 의 학습은 문장과 문장 간에 관계를 학습 parameter 가 representation 하기에 좋다
  - 현재 문장과 다음 문장 생성이 좋기는 하지만, Resource 가 많이…more
  - 따라서 다음 Sentence 를 예측하는 Pretraining 할 때
    - A -> B(50%), (random sentence 50%) 이용
  - 그리고 Output 에서 Not Next 인지 Is Next 인지를 학습함(binary classification)
    - 이를 통해서 문장과 문장 사이에 관계를 학습하는 Pre-training 시작
    - 이게 효과적이다…
  - 확실히 좋은 Performance 가 난다는 것을 증명

        - OpenAI-GPT 와 다른 점은 문장과 문장 사이에 구분자를 Pre-Training 과정에 학습하고, A 와 B 가 다른 문장이라는 것도 인지, 파인 튜닝에서 모델 파라미터를 전체 학습
    - 3.4 pre-training Procedure
        - Training 을 어떻게 할 것인가
        - LM 에 필요한 데이터 구성
        - A->B sentence (50% random) -> batch 당 128,000 개
        - TPU 사용
        - 4 일간 Training
    - 3.5 Fine-Tuning Procedure
        - Sequence-level classification tasks straight forward
        - Fine-tuning 은 classification layer 에서 사용
        - Fine-tuned jointly to maximize the log-probability
        - 대부분 pre-training model 은 same model 임 batch size, learning rate 와 training epochs 가 다름
        - 대부분의 fine-tuning 은 엄청 빨리 학습된다
    - 3.6 Comparison of BERT and OpenAI GPT
        - GPT 와 학습 데이터 및 학습 방법의 조금씩 차이가 있지만, chapter 5 ablation 실험에서 분명히 BERT 가 더 효과적으로 적용이 가능하다는 것을 보여준다. 직접보시기를…
- 4. Experiments
    - 4.1 GLUE(General Language Understanding Evaluation) Dataset
        - 일관적인 테스트 및 학습을 위해 데이터 셋 사용
        - MNLI(Multi-Genre Natural Language Inference)
            - 한 쌍의 문장이 주어지면, 첫 번째 문장에 비교하여 두 번째 문장이 함정, 대조, 중립인지 확인하는 것
        - QQP(Quora Question Pairs is a binary classification task)
            - 2 개의 질문이 같은 질문인지 아닌지 파악하는 것
        - QNLI(Question Natural Language Inference)
            - Stanford Question Answering
        - SST-2
            - Stanford Sentiment Treebank
            - Sentiment 분류
        - CoLA
            - Corpus of Linguistic Acceptability is binary single-sentence classification task
            - English sentence 가 Linguistically 한지
        - STS-B

- 2 개 Sentence 가 비슷한지 1~5 점

- MRPC
  - Microsoft Research Paraphrase Corpus Consists of Sentence pairs Automatically Extracted 되는지 온라인 뉴스 sentence 들이 Semantic 하게 동등한지 확인
- Total Figure



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

- RTE
  - Recognizing Textual Entailment(함정)
  - Like MNLI
  - Winograd NLI is a small natural language inference dataset deriving
- Result

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); $\text{BERT}_{\text{BASE}}$ = (L=12, H=768, A=12); $\text{BERT}_{\text{LARGE}}$ = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from https://gluebenchmark.com/leaderboard and https://blog.openai.com/language-unsupervised/.

- 4.1.1 GLUE Result
  - 모든 위에 Task 에서 BERT 결과가 좋다

- 4.2 SQuAD v1.1
  - Data set
    - Input Question:

      `Where do water droplets collide with ice crystals to form precipitation?`

    - Input Paragraph:

      `... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...`

    - Output Answer:

      `within a cloud`

    ○

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

  -
  - Human Score 보다 5%높다
- 4.3 NER

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

Table 3: CoNLL-2003 Named Entity Recognition results. The hyperparameters were selected using the Dev set, and the reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

- 4.4 SWAG
  - Situation with Adversarial Generation dataset contain
    - Ex)

A girl is going across a set of monkey bars. She

(i) jumps up across the monkey bars.

(ii) struggles onto the bars to grab her head.

(iii) gets to the end and stands on a wooden plank.

(iv) jumps up and does a back flip.

- 결과

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Table 4: SWAG Dev and Test accuracies. Test results were scored against the hidden labels by the SWAG authors. †Human performance is measure with 100 samples, as reported in the SWAG paper.

- 5. Ablation Studies
  - 음….
- 6. Conclusion
  - 학습 사이즈, 모델 크기, 그걸 돌릴 수 있는 Resource 가 있느냐..
  - Google !!
- 7. Next
  - 한국어로 구현…
  - 자원부족…
  - 어떻게? 고민
- 참조:
  - https://arxiv.org/abs/1810.04805

- https://rosinality.github.io/2018/10/bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding/
- https://www.facebook.com/hunkims/videos/10156797151174521/