



[8주차] 논문리뷰

BERT

1. Introduction

논문이 다루는 분야

해당 task에서 기존 연구 한계점

논문의 contributions

2. Related Work

2.1. Unsupervised Feature-based Approaches

2.2. Unsupervised Fine-tuning Approaches

2.3. Transfer Learning from Supervised Data

2.4. 기존 연구와 BERT의 차별성

3. 제안 방법론

Main Idea

Model Architecture

Input/Output Representations

Pre-training BERT

Fine-tuning BERT

Contribution

4. 실험 및 결과

4.1. GLUE

4.2. SQuAD v1.1

4.3. SQuAD v2.0

4.4. SWAG

5. Ablation Studies

5.1. Effect of Pre-training Tasks

5.2. Effect of Model Size

5.3. Feature-based Approach with BERT

6. 결론 (배운점)

ViT

1. Introduction

논문이 다루는 분야

해당 task에서 기존 연구 한계점

논문의 contributions

BERT

1. Introduction



논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가
논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

- **BERT**라는 새로운 언어 표현 모델을 제안
- BERT는 왼쪽과 오른쪽 문맥을 모두 고려하여
→ 딥 양방향 표현(deep bidirectional representations) 을 사전 학습(pre-train)하고
→ 이를 다양한 자연어처리(NLP) 작업에 쉽게 fine-tuning 할 수 있도록 설계됨.
- 기존 사전 학습된 언어 모델들은 주로 한 방향으로만 문맥을 고려
→ 문장의 양쪽 맥락을 동시에 고려할 수 없다는 한계가 有
→ 문장 수준과 토큰 수준 작업에서는 양방향 문맥 이해가 매우 중요

논문이 다루는 분야

- 자연어처리 (NLP)
- 언어 모델링(language modeling)
- 전이 학습(transfer learning)
- 텍스트 이해(text understanding)

해당 task에서 기존 연구 한계점

- feature-based 방법: 양방향 문맥을 concatenation
- fine-tuning 기반 방법: 왼쪽-오른쪽 방향성만 학습 / 문장 전체의 양방향 정보를 제대로 반영X

⇒ 기존 방법은 구조적으로 양방향 문맥을 깊게 고려하는 데 한계가 있었음.

논문의 contributions

- **Masked Language Model (MLM)** 을 통해 진정한 딥 양방향 사전학습을 가능하게 함
- 다양한 NLP 작업에서 복잡한 task-specific 구조 없이 fine-tuning만으로 최고 성능 달성
- **11개 NLP** 벤치마크에서 새로운 SOTA 기록 달성

2. Related Work



Introduction에서 언급한 기존 연구들에 대해 어떻게 서술하는가
제안 방법의 차별성을 어떻게 표현하고 있는가

2.1. Unsupervised Feature-based Approaches

- 전통적 단어 임베딩연구 (ex. Word2Vec, GloVe)
- **ELMo**: 왼쪽-오른쪽 별도 학습 후 결과를 결합하여 문맥 정보를 반영. 하지만 깊은 양방향 표현은 아님.

2.2. Unsupervised Fine-tuning Approaches

- **OpenAI GPT**: 좌측 맥락만 보는 Transformer로 전체 문장을 예측하는 방식.
- **ULMFiT**. fine-tuning으로 downstream task를 학습.

2.3. Transfer Learning from Supervised Data

- Supervised task(ex. 자연어 추론 데이터)를 이용하여 사전학습
- 컴퓨터 비전에서는 ImageNet pre-training 후 fine-tuning하는 접근법이 일반적.

2.4. 기존 연구와 BERT의 차별성

- BERT는 딥 양방향 사전학습이 가능하도록
→ Masked Language Model(MLM) 과 Next Sentence Prediction(NSP) 을 사용.
- 사전학습-미세조정(fine-tuning) 일관된 구조 사용.
- 단순히 feature를 추출하는 데 그치지 않고, 모든 파라미터를 통째로 **fine-tuning**.

3. 제안 방법론



Introduction에서 언급된 내용과 동일하게 작성되어 있는가
Introduction에서 언급한 제안 방법이 가지는 장점에 대한 근거가 있는가
제안 방법에 대한 설명이 구현 가능하도록 작성되어 있는가

Main Idea

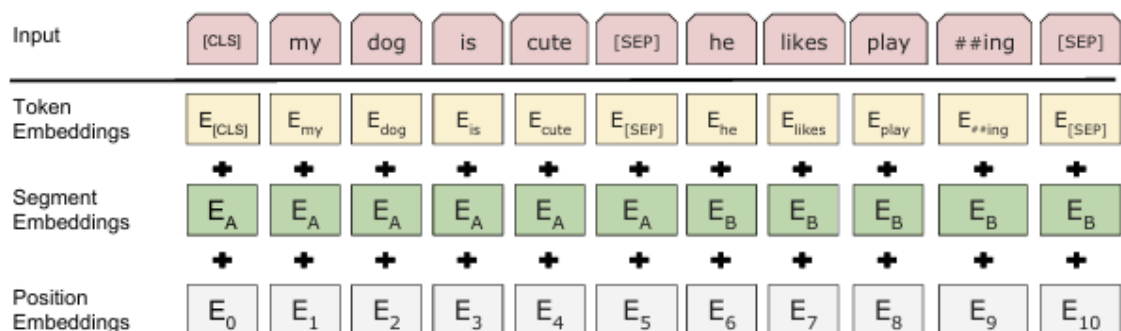
- 입력 문장을 WordPiece로 잘라서 token sequence로 변환
- special token CLS와 SEP를 추가.
- 입력은 단일 문장 또는 문장 쌍 모두 가능.

Model Architecture

- **Transformer Encoder** 구조 기반.
- 두 모델 구조
 - : BERTBASE (12-layer, 768-hidden, 12-head)
 - : BERTLARGE (24-layer, 1024-hidden, 16-head)

Input/Output Representations

- 입력 임베딩: Token Embedding + Segment Embedding + Position Embedding
- CLS 벡터는 전체 시퀀스를 대표하는 표현으로 사용



Pre-training BERT

Task 1: Masked LM

- 입력 토큰 중 15%를 무작위로 마스킹, 주변 문맥을 기반으로 해당 토큰 예측
- denoising autoencoder와 유사 BUT, 전체 문장이 아니라 마스킹된 부분만 예측

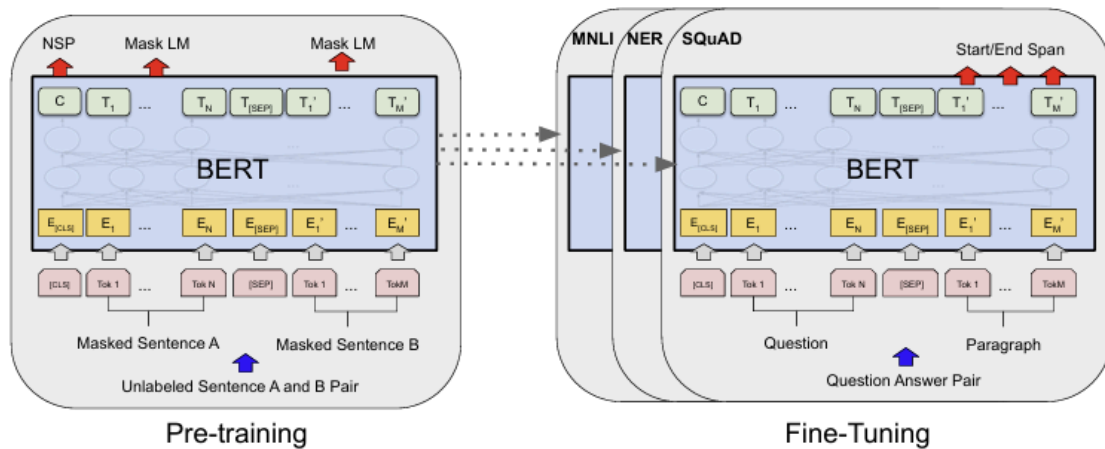
Task 2: Next Sentence Prediction (NSP)

- 문장 A 다음에 문장 B가 실제로 이어지는지 여부를 이진 분류(IsNext / NotNext)로 예측

Fine-tuning BERT

- 사전학습된 BERT에 task-specific output layer만 추가.

- 모든 downstream 작업(text classification, QA 등)에서 같은 구조 사용.



Contribution

- Masked Language Model (MLM) 사용하여 deep bidirectional pre-training 을 가능하게 함
- 문장 쌍 간의 관계 학습을 모델링하기 위해 Next Sentence Prediction (NSP) 도입
- Pre-training과 Fine-tuning 과정에서 구조 일관성 유지 → 전이학습(transfer learning) 단순화
 - Task-specific output layer만 추가, 전체 모델을 end-to-end로 미세조정 가능하게 함
- 다양한 Downstream Task를 하나의 통합 모델로 처리
- Feature-based 접근과 Fine-tuning 접근 모두에서 뛰어난 효과
 - 전체 모델을 fine-tuning 하는 방식 외에도, 특정 hidden layer feature를 뽑아서 별도 task-specific 모델에 활용해도 높은 성능을 보임.

4. 실험 및 결과



Introduction에서 언급한 제안 방법의 장점을 검증하기 위한 실험이 있는가

4.1. GLUE

- 다양한 문장 수준/문장 쌍 수준 NLU tasks 집합.

- BERTBASE와 BERTLARGE 모두 GLUE에서 기존 모델보다 높은 평균 정확도를 기록함.
- Dataset
 - GLUE benchmark (문장 분류, 문장 쌍 분류 등 다양한 NLU 과제 모음)
 - MNLI, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, RTE 포함
- Baseline
 - Pre-OpenAI SOTA 모델 (BiLSTM+ELMo 등)
 - OpenAI GPT
- 결과
 - BERTBASE: 기존 최고 모델(OpenAI GPT) 대비 평균 +4.5% 성능 향상
 - BERTLARGE: 평균 +7.0% 성능 향상
 - 특히, MNLI task에서 +4.6% 절대 정확도 개선
 - GLUE leaderboard 기준 BERTLARGE: 80.5점 → OpenAI GPT 대비 대폭 향상

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	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

4.2. SQuAD v1.1

- 질문과 단락을 받아 정답이 포함된 스패를 예측.
- BERT가 SQuAD v1.1 단일 모델 기준 최고 성능을 기록함.
- Dataset
 - 100k개 이상의 crowd-sourced 질문-답변 쌍 (Wikipedia 기반)
- Baseline
 - BiDAF+ELMo (Single model)
 - R.M. Reader (Ensemble)
- 결과

- BERTLARGE (Single model): F1 91.1% (TriviaQA 없이도 90.9%)
- BERTLARGE (Ensemble): F1 93.2% → 기존 최상위 시스템 대비 +1.5 F1 개선
- Single model만으로도 기존 SQuAD leaderboard ensemble을 초과

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT _{BASE} (Single)	80.8	88.5	-	-
BERT _{LARGE} (Single)	84.1	90.9	-	-
BERT _{LARGE} (Ensemble)	85.8	91.8	-	-
BERT _{LARGE} (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERT _{LARGE} (Ens.+TriviaQA)	86.2	92.2	87.4	93.2

4.3. SQuAD v2.0

- 답이 없는 경우도 허용하는 SQuAD 확장 버전.
- no-answer를 CLS 토큰 위치로 모델링하여, 답이 없을 때를 잘 분류할 수 있도록 학습.
- Dataset
 - 답변이 존재하지 않는 경우를 포함한 질문-답변 쌍
- Baseline
 - MIR-MRC (F-Net)
 - nlnet
 - Human upper bound도 비교
- 결과
 - BERTLARGE (Single model):
 - Dev set: EM 80.0, F1 83.1
 - Test set: EM 80.0, F1 83.1
 - 이전 최고 성능 대비 +5.1 F1 개선

- 인간 성능(약 89.5 F1)에는 미치지 못하지만, 기존 모델들 대비 큰 도약

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-	-	71.4	74.4
Ours				
BERT _{LARGE} (Single)	78.7	81.9	80.0	83.1

4.4. SWAG

- 주어진 문장에 가장 자연스러운 다음 문장을 선택하는 상식추론 데이터셋.
- BERT가 SWAG에서도 기존 모델 대비 크게 성능 향상.
- Dataset
 - 113k개 sentence-pair completion 예제 (상식적 추론 문제)
- Baseline
 - ESIM+ELMo
 - OpenAI GPT
- 결과
 - BERT_{LARGE}:
 - Dev accuracy: 86.6%
 - Test accuracy: 86.3%
 - OpenAI GPT 대비 +8.3% 향상
 - 기존 최고 baseline (ESIM+ELMo) 대비 +27.1% 절대 향상

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
BERT _{BASE}	81.6	-
BERT _{LARGE}	86.6	86.3
Human (expert) [†]	-	85.0
Human (5 annotations) [†]	-	88.0

5. Ablation Studies

- Ablation Study는 모델의 성능에 가장 큰 영향을 미치는 요소를 찾기 위해
- 모델의 구성요소 및 feature들을 단계적으로 제거 하거나 변경해가며 성능의 변화를 관찰하는 방법
- 모델의 핵심적인 구성요소와 하이퍼파라미터등을 파악할 수 있습니다.

<https://modulabs.co.kr/blog/ablation-study>.

5.1. Effect of Pre-training Tasks

- NSP를 제거하거나, 왼쪽-오른쪽 방향만 학습하면 성능이 상당히 하락
- 특히 QNLI, MNLI, SQuAD와 같은 문장 관계 추론 작업에서 NSP의 중요성이 확인됨.
- Dataset
 - MNLI, QNLI, MRPC, SST-2, SQuAD v1.1 (BERT 논문에서 실험한 대표적 downstream tasks)
- Baseline
 - Full BERTBASE (MLM + NSP 적용)
 - 변형된 모델들:
 - No NSP (Next Sentence Prediction 제거)
 - LTR & No NSP (Left-to-Right만 학습, NSP도 제거 — OpenAI GPT에 가까운 설정)
 - 추가로 BiLSTM을 LTR모델 위에 얹은 버전도 실험
- 결과

- No NSP:
 - 특히 문장쌍 관계(task)에서 성능 하락 (ex. QNLI, MNLI).
 - SQuAD에서도 성능 약간 감소.
- LTR & No NSP:
 - 모든 task에서 큰 성능 저하 발생.
 - 특히 MRPC(문장 유사성)와 SQuAD(질문응답)에서 성능 급락.
- LTR & No NSP + BiLSTM 추가:
 - 일부 성능 회복했지만 여전히 양방향 사전학습 성능에는 미치지 못함.

Tasks	Dev Set				
	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
BERT _{BASE}	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

5.2. Effect of Model Size

- 더 큰 모델(BERT_{LARGE})이 항상 더 높은 성능을 보임.
- 작은 데이터셋(MRPC)에서도 큰 모델의 이점이 뚜렷하게 나타남.
- Dataset
 - MNLI, QNLI, MRPC, SST-2 (대표적인 GLUE 과제)
- Baseline
 - BERT 모델 크기 변화 실험:
 - (L=3, H=768, A=12)
 - (L=6, H=768, A=3 또는 12)
 - (L=12, H=768, A=12) → BERT_{BASE} 수준
 - (L=12, H=1024, A=16)
 - (L=24, H=1024, A=16) → BERT_{LARGE} 수준
- 결과
 - 모델 크기(L, H, A)가 커질수록 모든 task에서 성능 지속적 증가.

- **MNLI, MRPC** (데이터 수 적은 task)에서도 큰 모델이 더 좋은 성능.
- Language modeling perplexity (LM ppl)도 모델이 클수록 낮아짐.

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

5.3. Feature-based Approach with BERT

- Fine-tuning 없이 BERT의 hidden layer 출력을 feature로 활용할 경우에도 매우 좋은 결과.
- 특히 마지막 네 개 hidden layer를 concat하는 방식이 가장 우수.
- **Dataset**
 - CoNLL-2003 Named Entity Recognition (NER) Task
- **Baseline**
 - Fine-tuning 방식 (BERT 전체 미세조정)
 - Feature-based 방식 (BERT의 특정 layer feature만 뽑아서 사용)
- **결과**
 - Fine-tuning (BERTLARGE): Test F1 92.8
 - Feature-based (BERTBASE 기준):
 - Single layer 사용: 약간 낮은 성능 (ex. Last hidden 94.9 Dev F1)
 - Top 4 hidden layers concat: Dev F1 96.1 (Fine-tuning 대비 0.3점 차이)
 - 전체적으로, feature extraction만으로도 기존 SOTA 모델들과 비슷하거나 더 좋은 성능.

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	93.1
Fine-tuning approach		
BERT _{LARGE}	96.6	92.8
BERT _{BASE}	96.4	92.4
Feature-based approach (BERT _{BASE})		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

6. 결론 (배운점)



연구의 의의 및 한계점, 본인이 생각하는 좋았던/아쉬웠던 점 (배운점)

- 언어 모델 기반 전이학습이 NLP 전반에 중요함.
- 딥 양방향 Transformer를 통한 사전학습이 매우 효과적임 (BERT)
- 다양한 NLP task에 대해 별도의 아키텍처 수정 없이 뛰어난 성능을 발휘할 수 있음.

ViT

1. Introduction



논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가
 논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

- 기존 CNN 중심의 컴퓨터 비전 구조를 벗어나, 표준 Transformer를 이미지에 직접 적용하여 이미지 분류 작업을 수행하는 방법을 다룸.

- 구체적으로는 이미지를 16×16 패치 단위로 나누고, 이를 토큰처럼 다뤄서 Transformer에 입력하는 방식을 제안함.
- NLP에서는 Transformer 기반 모델들이 대규모 학습을 통해 성능이 계속 향상되고 있는 반면 컴퓨터 비전에서는 여전히 CNN이 지배적이고, Transformer를 적용하는 연구는 제한적.
- Transformer가 충분한 데이터만 주어진다면 CNN의 구조적 편향 없이도 뛰어난 성능을 낼 수 있는지를 실험

논문이 다루는 분야

- 컴퓨터 비전(Computer Vision)
- 이미지 분류(Image Classification)
- 전이 학습(Transfer Learning), 대규모 사전학습(Pretraining at Scale)

해당 task에서 기존 연구 한계점

- Transformer를 비전에 적용하는 기존 연구들은 대부분 CNN 기반 구조를 유지하거나, 특수한 어텐션 형태를 사용해 왔음.
 - 하드웨어 가속기에서 비효율적
 - 모델을 대규모로 확장하기 어려움
 - NLP처럼 간단하고 일관된 확장 전략X
- 또한 CNN은 구조적으로 귀납적 편향이 강해서 충분한 데이터 없이도 잘 작동하지만, Transformer는 데이터가 적을 경우 일반화 성능이 떨어진다는 단점이 있음.

논문의 contributions

1. 순수한 Transformer만으로 이미지를 직접 처리하는 모델을 제안
2. 이미지를 패치 단위로 나누어, 텍스트의 토큰처럼 취급하는 방법을 설계
3. 중간 규모 데이터셋(ImageNet)에서는 ResNet 대비 소폭 낮은 성능을 보이지만, 대규모 데이터셋(14M~300M 이미지)에서는 CNN 기반 모델을 능가하는 성능을 달성
4. 사전학습 규모가 충분히 크면, Transformer도 CNN의 귀납적 편향 없이 강력한 일반화 성능을 보일 수 있음을 입증
5. 공개된 ImageNet-21k나 사내 JFT-300M 데이터셋을 활용하여, 다양한 이미지 분류 벤치마크에서 최첨단(SOTA) 또는 그 이상의 성능을 달성함.