What Does BERT Look At?
An analysis of BERT's Attention

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

https://www.youtube.com/watch?v=jT4KufIqM_E

2022. 05. 11

목차

I. Introduction
II. Background: Transformers and BERT
III. Experiments
    i. Surface-Level Patterns in Attention
    ii. Probing Individual Attention Heads
    iii. Clustering Attention Heads
IV. Conclusion

1. Introduction

- 모델이 왜 그렇게 결정을 하는가, 모델이 무엇을 학습하고 있는가 등 XAI에 관심
- BERT 모델은 아직까지도 좋은 성능의 모델로 사용되고 있고, BERT를 기반으로 다양한 테크닉을 추가한 좋은 성능의 모델들이 개발되고 있음.

BERT는 왜 성능이 좋은가?

- 모델이 사전학습을 통해 언어 구조를 파악함 (주어, 동사 등)

정말 언어적 특징을 학습하는 것인지 어떻게 확인할 수 있는가?

- 선별된 문장을 입력하여 결과를 통해 확인
- 중간 산출물인 vector representation 으로 probing classifier 를 평가

본 논문 에서는 BERT 모델의 Attention maps을 통해 각 attention head의 특성을 확인함.

1. Introduction

1. **전반적인 Attention head 의 동작 분석 (Surface-Level Patterns in Attention)**
   - 대부분의 attention head가 특정 토큰에 집중 (CLS, SEP, next 등)
   - [SEP] 토큰의 역할
2. **개별 Attention head 의 동작 분석 (Probing Individual Attention Heads)**
   - Dependency parsing
   - Coreference resolution
3. **Attention heads 의 군집분석 (Clustering Attention Heads)**
   - 비슷한 위치의 레이어의 attention head 들의 동작이 비슷함.

2. Background – Transformers and BERT

Transformers 란?

attention을 활용한 encoder-decoder 구조의 모델

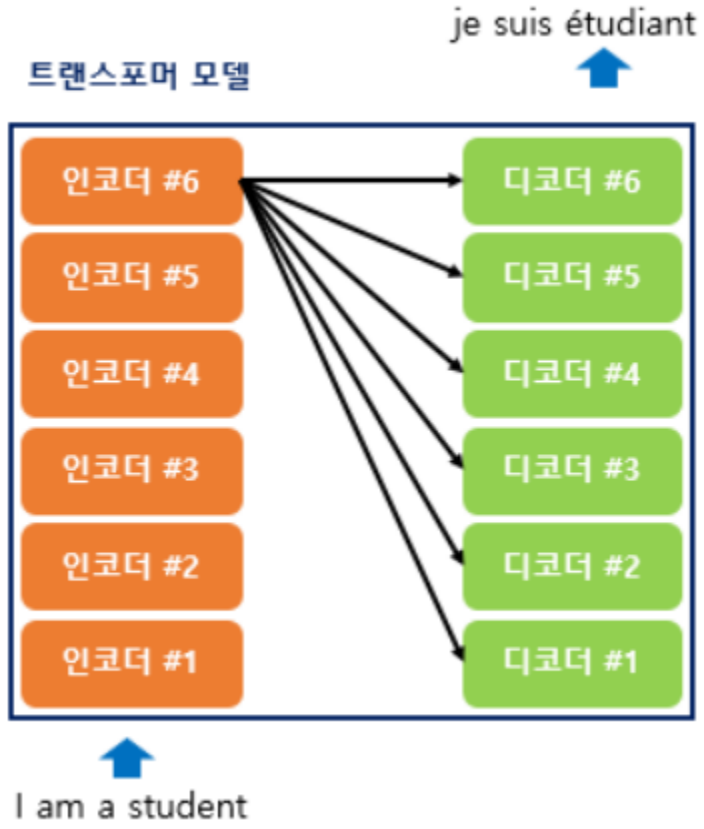
$$\text{Attention}(X, Y, Z) = \text{softmax}\left(\frac{\text{mask}(XY^\top)}{\sqrt{d}}\right)Z$$

$$\text{MultiHead}(X, k) = [h_1; \cdots; h_k]W_o$$

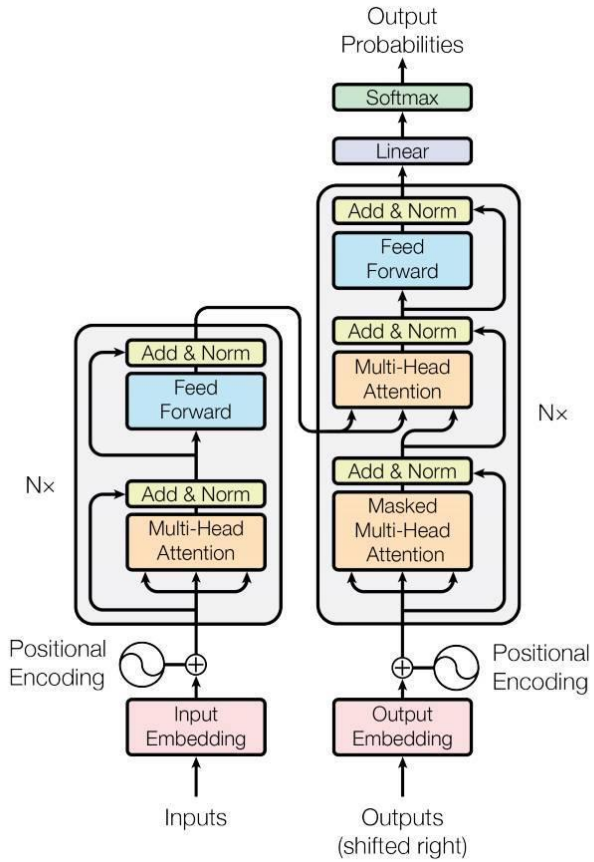
$$\text{where } h_j = \text{Attention}(XW_j^1, XW_j^2, XW_j^3)$$

attention: Q, K, V 로 단어들간
연관성을 구함

Multi-head Attention:
여러 번의 attention을 병렬로 사용



https://wikidocs.net/31379

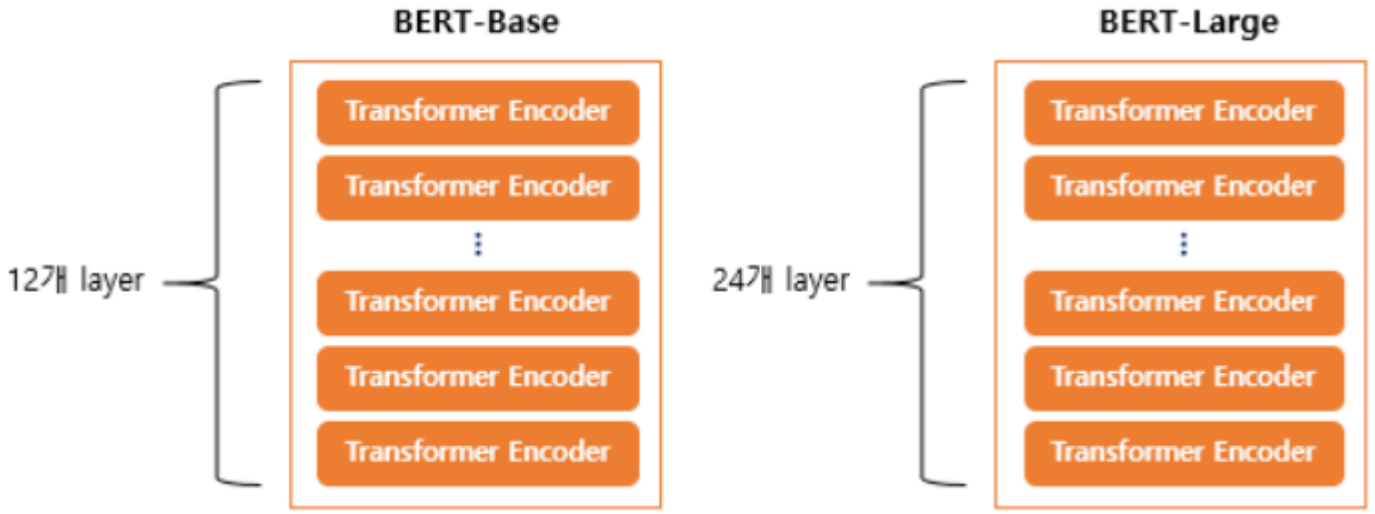


Transformer architecture

2. Background – Transformers and BERT

BERT 란?

Transformer의 encoder를 여러 층으로 쌓아 올린 구조의 모델

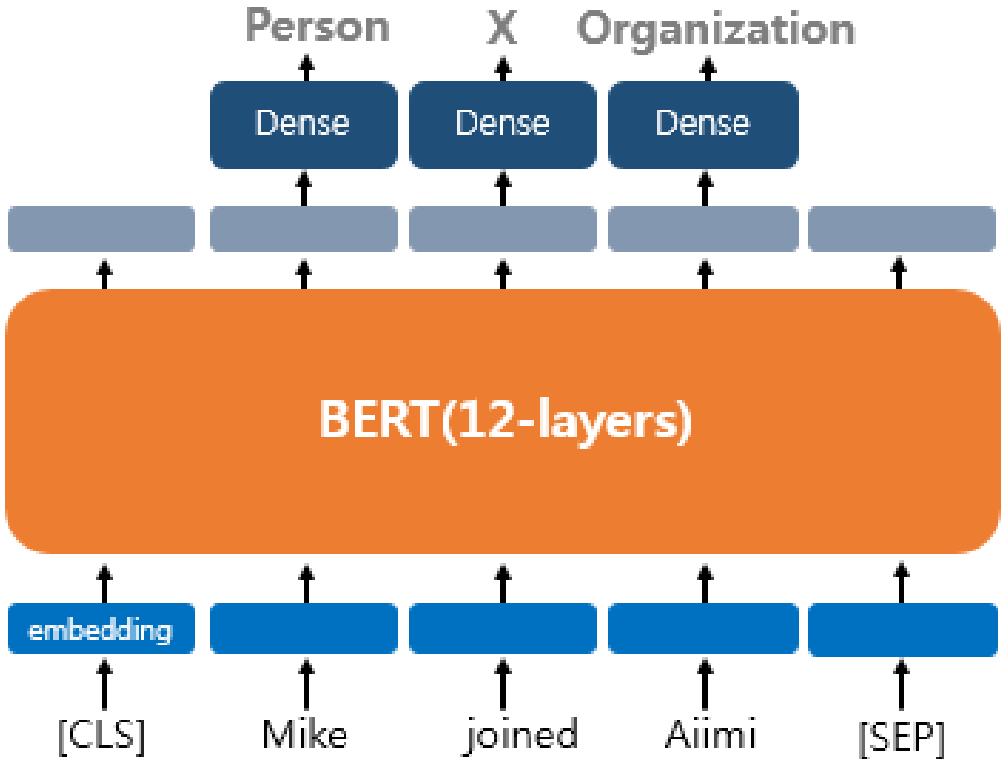


2. Background – Transformers and BERT

BERT 란?

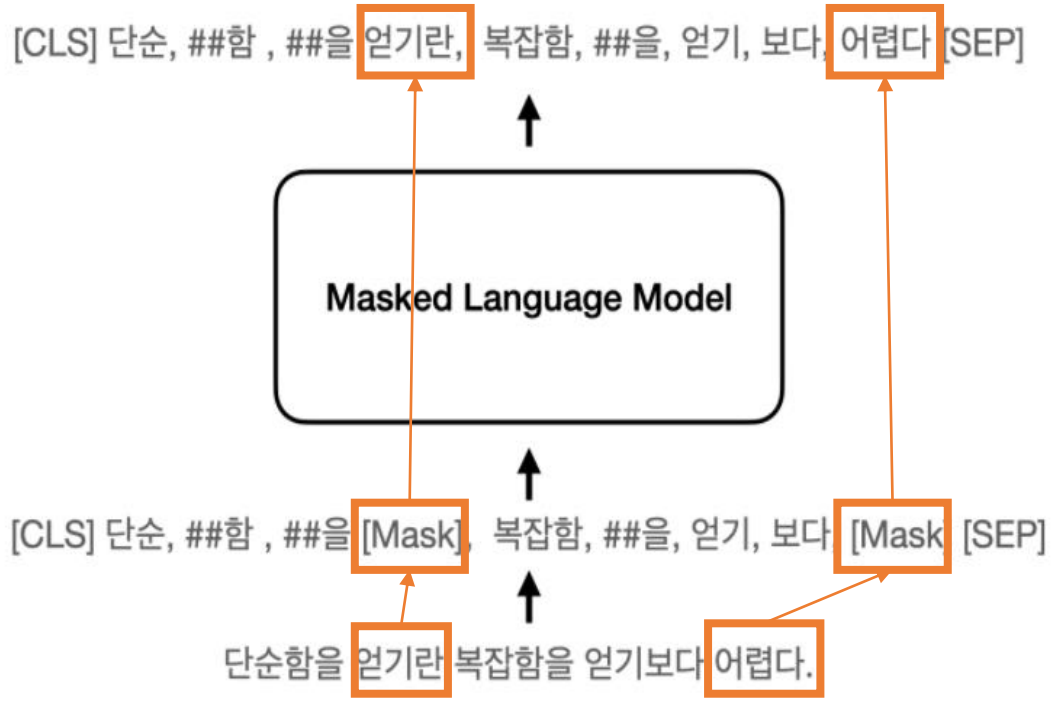
각 layer 마다 multi-head attention 과 feed-forward NN 을 수행

모든 단어들을 참고하여 문맥을 반영한 output embedding 을 얻음



2. Background – Transformers and BERT

BERT의 사전학습 방식



MLM(Masked Language Modeling)

: 입력으로 사용하는 문장의 토큰 중 15%의 확률로 선택된 토큰을 [MASK] 토큰으로 변환시키고, 언어모델을 통해 변환되기전 [MASK] 토큰을 예측하는 언어모델링 방법으로 학습

3. Experiments 1) Surface-Level Patterns in Attention

Attention head 가 어떻게 동작하는지 표면 수준 패턴 분석 수행 (일반적인 행동)

Setup.

- Input: [CLS] 문단1 [SEP] 문단2 [SEP]

  위키피디아의 1000개의 문단 사용,

  연속된 두 문단에 해당하는 최대 128개의 토큰을 입력으로 사용

- 입력시 masking 은 사용 안 함 : attention 의 온전한 동작을 보기 위함.
- Model 의 configuration 은 BERT-base 와 동일.

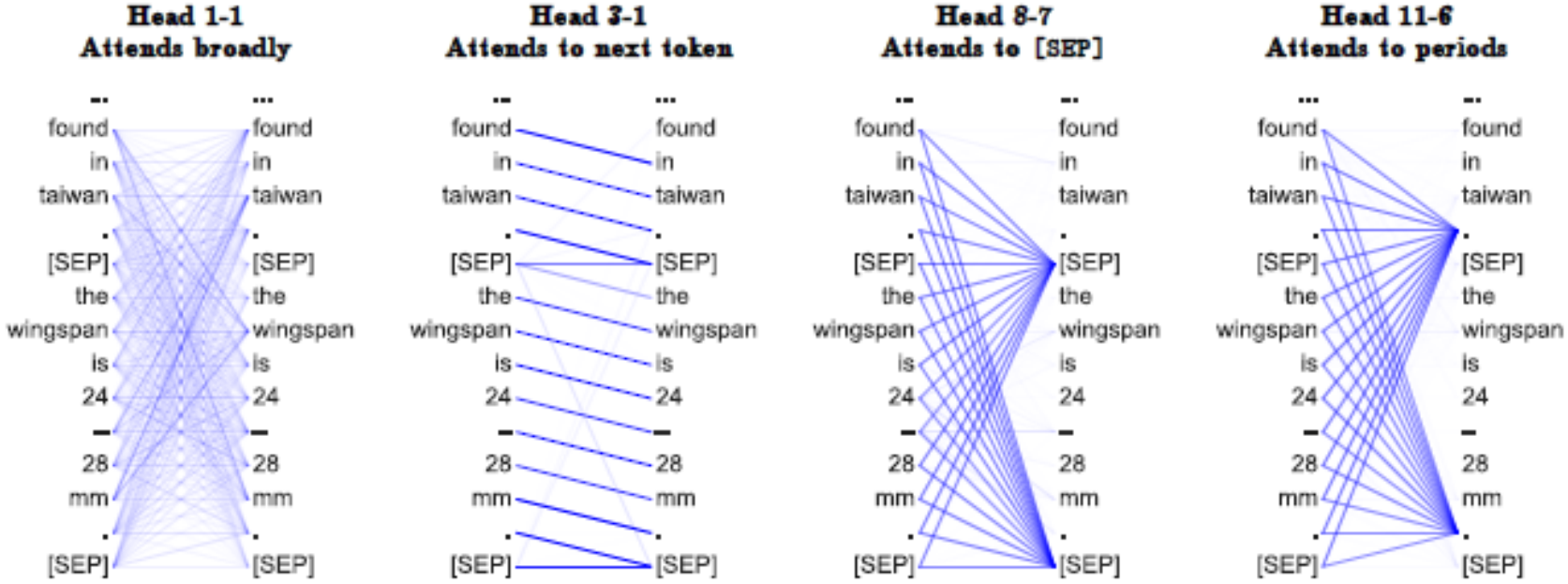
  12개의 layer, 768개의 hidden layer, 12개의 attention head, batch size = 16

3. Experiments 1) Surface-Level Patterns in Attention

Attention head 가 어떻게 동작하는지 표면 수준 패턴 분석 수행 (일반적인 행동)

Relative Position

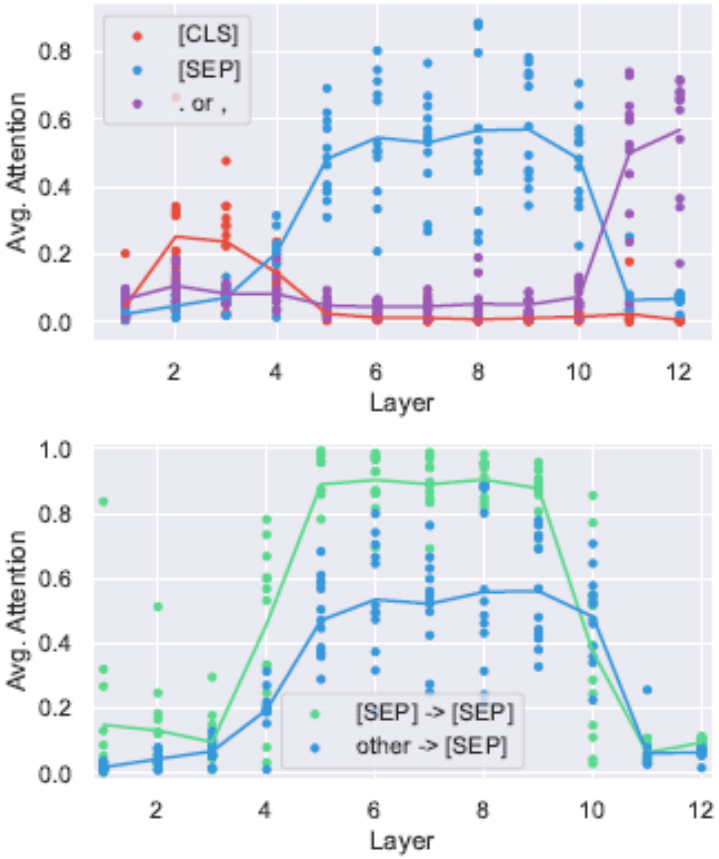
- BERT의 Attention 헤드가 현재 토큰, 이전 토큰 또는 다음 토큰에 얼마나 자주 관여하는지 계산한다
- 대부분의 헤드가 현재 토큰에 거의 주의를 기울이지 않는다는 것 발견
- 특히 next token 또는 previous token 에 많은 주의를 기울이는 head 들이 있었음
  - In layer 2, 4, 7, 8: 평균적으로 50% 이상을 'previous token'에 집중
  - In layer 1, 2, 2, 3, 6: 평균적으로 50% 이상을 'next token'에 집중



3. Experiments 1) Surface-Level Patterns in Attention

Attention head 가 어떻게 동작하는지 표면 수준 패턴 분석 수행 (일반적인 행동)

Attending to Separator tokens



- BERT가 몇 개의 토큰에 상당히 집중한다는 것 발견.
- CLS, SEP, '.', " 등의 토큰에 상당히 집중.
- 6-10 layer 에서 BERT의 관심 절반 이상이 [SEP] 에 집중
=> input data 에 이 토큰들이 항상 포함 되어있기 때문일 것이라고 추측

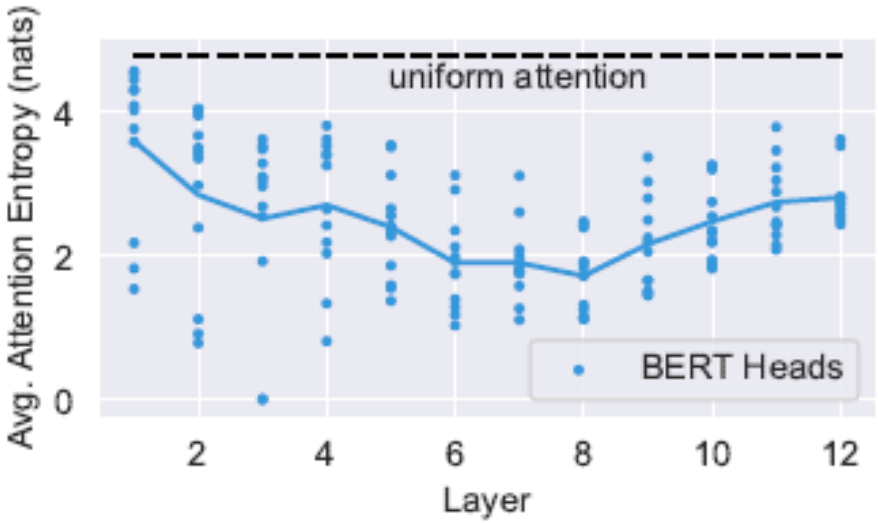
- 어떤 attention head 들에서는 동사-목적어, 전치사-명사 등의 관계에 attention 이 크게 걸리는 현상들이 나타나는데, 이 때 관련 없는 토큰들은 [SEP] 이 걸리는 현상을 보임.

[SEP] 의 역할
- MLM task 에서 attention 의 변화가 loss 값에 미치는 영향을 측정했을 때, [SEP] 토큰에 걸리는 attention 의 영향은 크지 않았음.
- 따라서 [SEP] 에 걸린 attention 들은 no-op (역할 없음)을 의미한다.

3. Experiments 1) Surface-Level Patterns in Attention

Attention head 가 어떻게 동작하는지 표면 수준 패턴 분석 수행 (일반적인 행동)

Focused vs Broad Attention



- Attention head 가 몇 개의 단어에 집중하는지 또는 많은 단어에 걸쳐 광범위하게 집중하는지 측정한다.
이를 위해 각 head 의 attention 분포의 평균 entropy 계산
(entropy ↑: 광범위한 attention, entropy ↓: 특정 단어에 attention
- [CLS] 토큰에서만 모든 attention head 에 대한 엔트로피 측정

- 마지막 레이어에서 [CLS] 토큰은 매우 광범위한 attention 값들을 갖고 있는데, CLS 토큰을 이용해 next sentence prediction 을 수행하므로 이러한 결과가 나온 것 같다.

3. Experiments 2) Probing Individual Attention Heads

Probing task 란?

- 모델이 어떤 언어적 정보를 파악할 수 있는지 확인하는 것

- 여러 가지 probing task 가 있음.

- 본 논문에서는 Dependency parsing, coreference resolution 수행

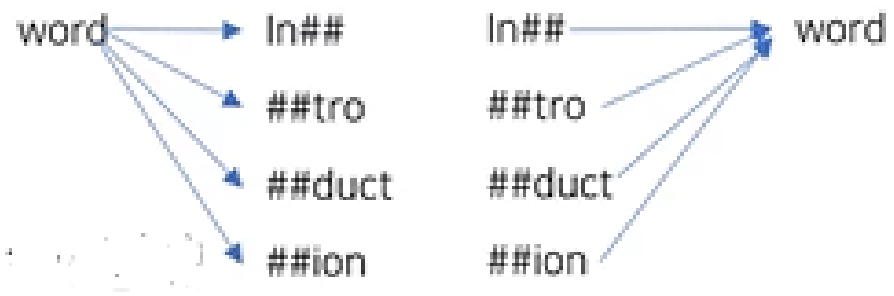
Dependency parsing: 의존하는 관계 확인 ex. 동사-목적어, 전치사-명사 등

Coreference resolution: 고유명사, 인칭대명사 등으로 지칭된 단어가 같은 것을 지칭하는지 확인

ex. She always studies hard. Her books are always old.

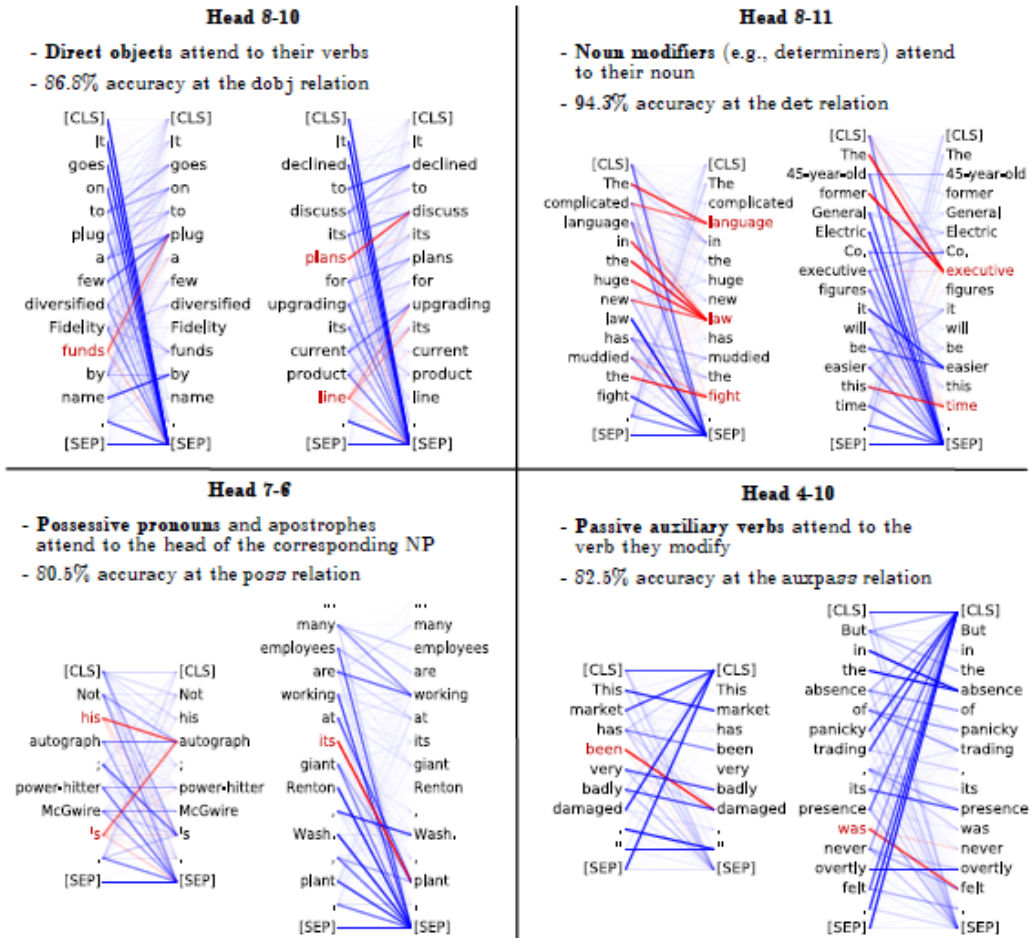
Words-level tasks 수행을 위해 token-token attention map 을 word-word 로 바꿈.

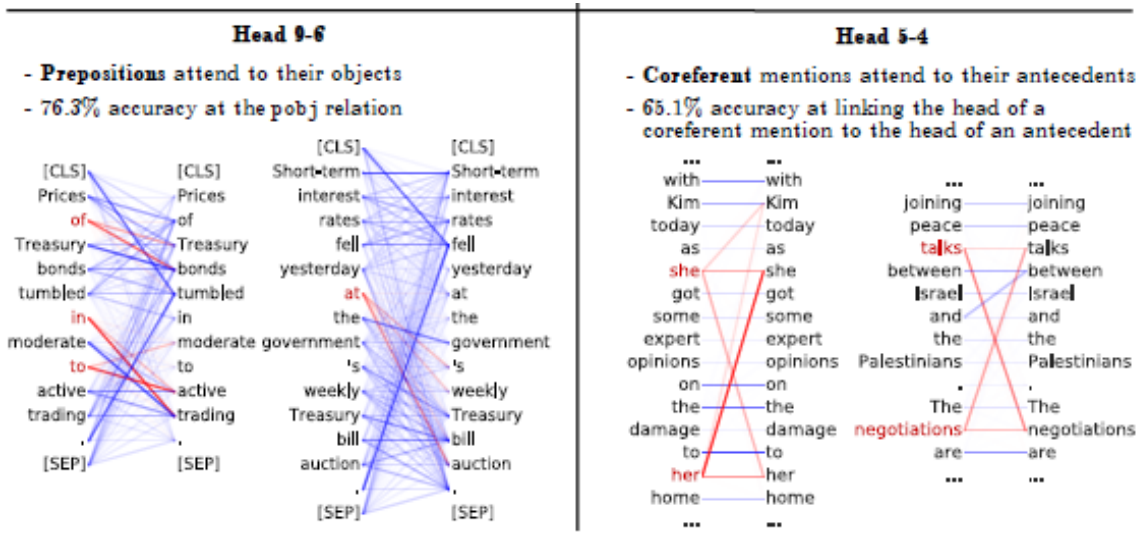
하나의 word 가 여러 개의 token에 영향을 준 경우: attention 을 합함



여러 개의 token이 하나의 word에 영향을 준 경우: attention의 평균을 구함

3. Experiments 2) Probing Individual Attention Heads





1. 목적어가 자신과 관련된 동사에 attention
   : funds -> plug
2. 관사 등(명사 결정자)이 자신이 가리키는 명사에 attention : The -> language
3. 소유격이 자신과 관련된 명사에 attention
   : its -> plant
4. 보조 동사가 동사에 attention
   : was -> felt
5. 전치사가 자신이 가리키는 명사에 attention
   : of -> bounds, to -> active
6. 이전에 언급된 다른 언급 참조
   : her -> she, negotiations -> talks

3. Experiments 2) Probing Individual Attention Heads

Relation	Head	Accuracy	Baseline
All	7-6	34.5	26.3 (1)
prep	7-4	66.7	61.8 (-1)
pobj	9-6	**76.3**	34.6 (-2)
det	8-11	**94.3**	51.7 (1)
nn	4-10	70.4	70.2 (1)
nsubj	8-2	58.5	45.5 (1)
amod	4-10	75.6	68.3 (1)
dobj	8-10	**86.8**	40.0 (-2)
advmod	7-6	48.8	40.2 (1)
aux	4-10	81.1	71.5 (1)
poss	7-6	**80.5**	47.7 (1)
auxpass	4-10	**82.5**	40.5 (1)
ccomp	8-1	**48.8**	12.4 (-2)
mark	8-2	**50.7**	14.5 (2)
prt	6-7	**99.1**	91.4 (-1)

- 모든 관계를 잘 포착하는 head는 없었지만,
특정 관계를 잘 포착하는 head 들이 있었음.

특정 head가 (다른 head 들에 비해)특히 잘하는 5가지 관계

3. Experiments 2) Probing Individual Attention Heads

Coreference Resolution

인칭대명사, 고유명사 등으로 나타낸 단어 중 같은 entity 를 지칭하는 단어를 찾아내는 task
Coreference resolution: 고유명사, 인칭대명사 등으로 지칭된 단어가 같은 것을 지칭하는지 확인

Model	All	Pronoun	Proper	Nominal
Nearest	27	29	29	19
Head match	52	47	67	40
Rule-based	69	70	77	60
Neural coref	83*	–	–	–
Head 5-4	65	64	73	58

*잘리지 않은 문서와 다른 언급 탐지를 통해 대략적으로만 비교할 수 있습니다.

- Nearest : 처음 명사가 등장하고 가장 가까운 mention 을 고름
- Head match: head word 가 매치되는지 확인. (head word: 동일한 키워드를 언급하는지 확인)
- Rule-based: 규칙기반 시스템
  전체 문자열 매치 확인 -> head word 매치 확인 -> 수, 성별, 인칭 매치 확인 -> 기타 mention 확인

- Neural coref: Sam Joshua Wiseman, Alexander Matthew Rush, Stuart Merrill Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In ACL.

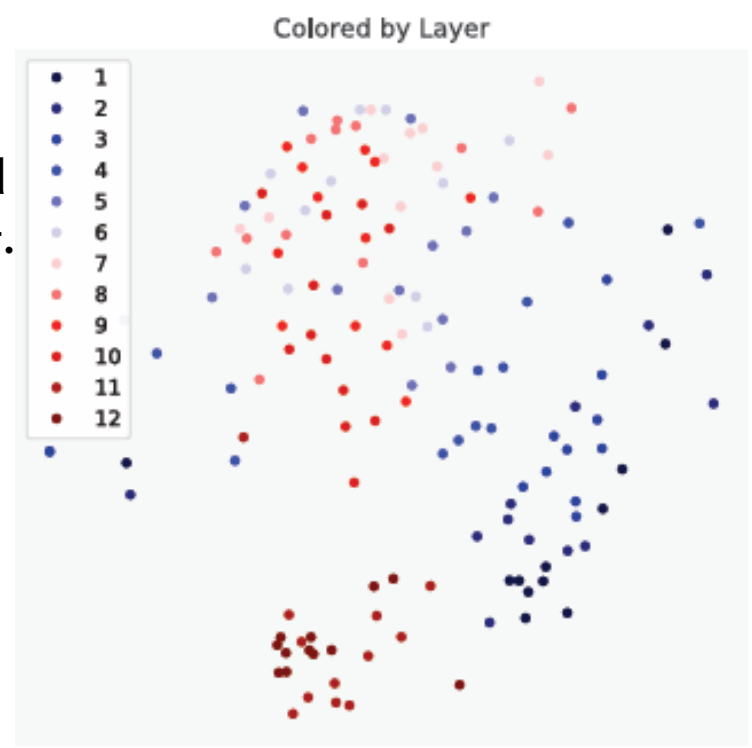
3. Experiments 3) Clustering Attention Heads

Clustering Attention heads

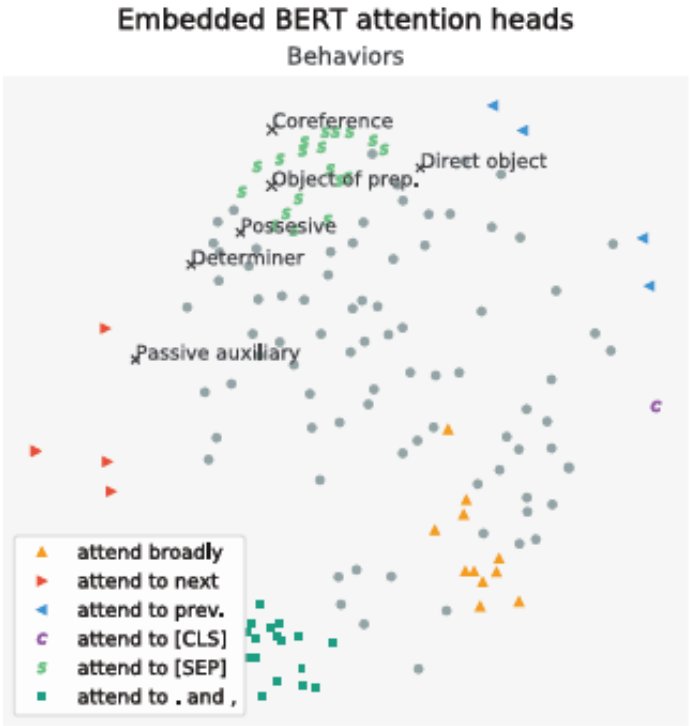
Jensen-Shannon Divergence 를 기준으로 클러스터링 진행

$$\sum_{\text{token} \in \text{data}} JS(\mathrm{H}_i(\text{token}), \mathrm{H}_j(\text{token}))$$

같은 레이어의 attention head 들끼리 모여 있음.



어떤 토큰에 집중하는 경향이 있는지에 따라 분포 확인



4. Conclusion

- NLP 모델 분석에 대한 연구는 모델 출력을 탐색하는데 초점을 맞춤.
- 본 연구에서는 attention map 에 초점을 맞추고 분석함.
- BERT 의 attention map 을 분석하여 BERT 가 상당한 언어 지식을 학습하고 있음을 실험을 통해 확인함.
- BERT 는 문장 구조를 어느 정도 학습하고 있음 (동사-목적어 등의 관계)

Thank you

What Does BERT Look At?
An analysis of BERT's Attention