Word Embedding

Winter Vacation Capstone Study

TEAM Kai.Lib

발표자 : 배세영

2020.02.24 (MON)

ELMo and BERT

- 언어 모델의 성능 향상을 위하여 모델의 가장 핵심이라 할 수 있는 워드 임베딩(Word Embedding) 방식을 개선하려는 시도가 있어 옴

- 2018년, ELMo와 BERT가 제안되며 NLP(Natural Language Processing) 분야의 트렌드를 주도함

기존 Word Embedding 방식의 문제점

- 주변 맥락 단어를 학습할 때만 고려하고, 이렇게 생성된 어휘 임베딩을 다른 모델의 입력으로 사용하는 상황은 가정하지 않았음

- 즉, 학습이 완료된 후 어휘 임베딩 값은 불변

- 실제로 사용되는 어휘의 의미는 맥락에 따라 가변적임
  - "나는 머리를 끄덕였다."
  - "나는 머리를 다시 잘랐다."
  - "나는 머리가 좋아서 공부를 잘 한다."

- 기존 임베딩 방식은 위의 세 가지 '머리' 단어에 대하여 같은 임베딩 값을 할당

- 학습이 잘 이루어졌다면 위의 세 가지 상황에 대한 맥락적 정보를 모두 포괄할 것이나 실제 사용되는 맥락과 관계 없는 정보까지 담고 있어야 하며, 이 정보는 사실상 불필요함

ELMo(Embedding from Language Models)

Deep contextualized word representations

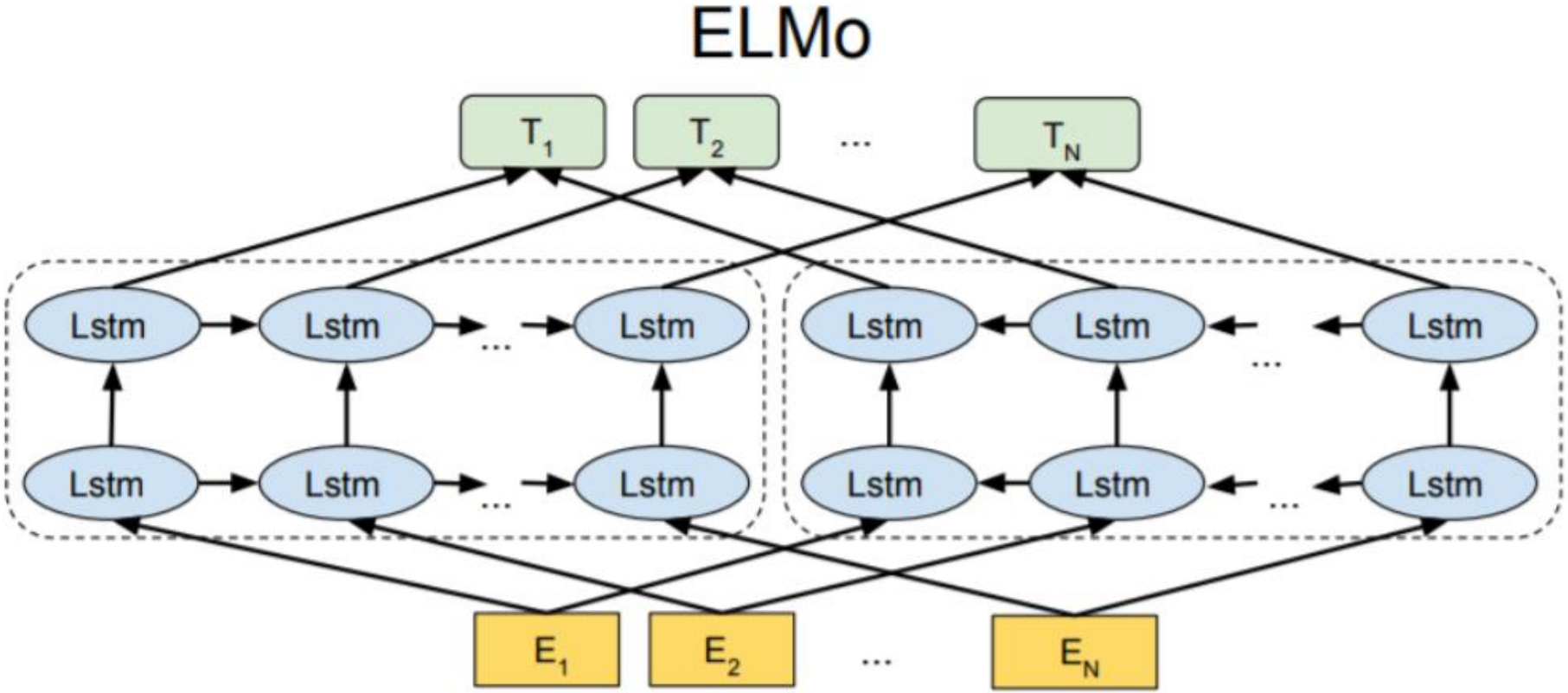
Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

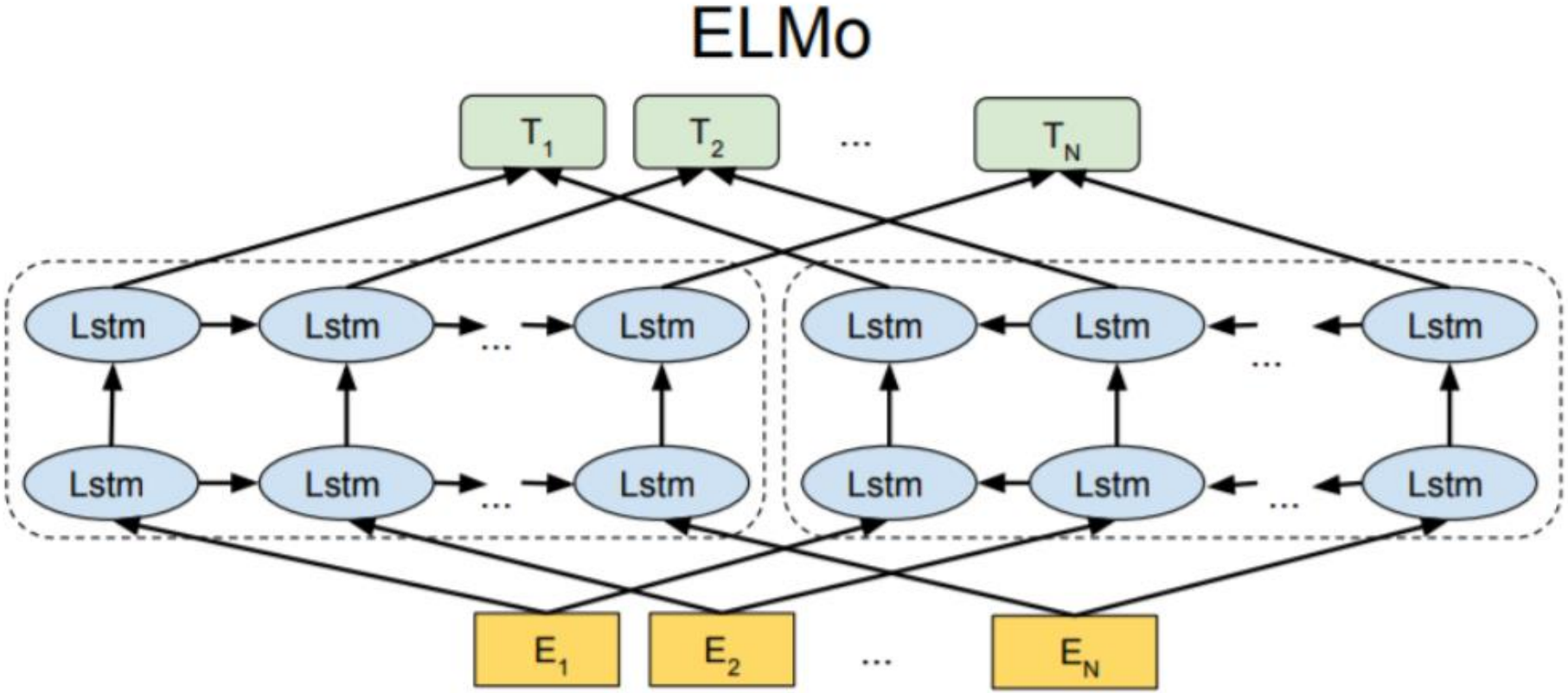
ELMo(Embedding from Language Models)

- 해당 단어의 Embedding Vector가 주변 단어로부터 추론된 맥락 정보에 따라 가변적일 수 있도록 함
- "맥락화된 어휘 임베딩(Contextualized Word Embedding)"



ELMo(Embedding from Language Models)

- 기존 Word Embedding 방식은 uni-directional함
- 정방향/역방향의 두 가지 uni-directional LSTM을 사용함으로써 단방향 embedding시 놓치고 지나가는 맥락 정보를 고려



ELMo(Embedding from Language Models)

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

BERT(Bidirectional Encoder Representations from Transformers)

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

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

BERT(Bidirectional Encoder Representations from Transformers)

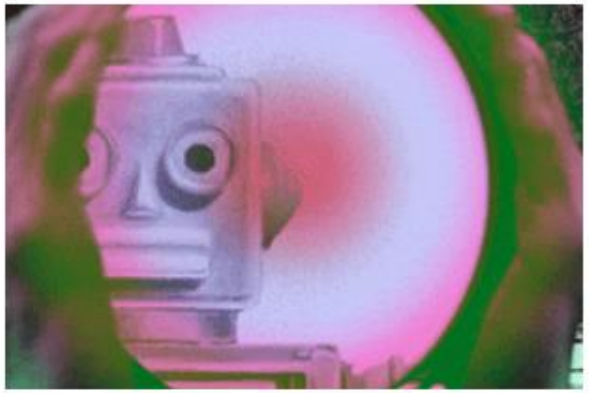
- 2018년 10월 논문 공개, 11월 오픈소스 공개로 혜성처럼 등장한 구글의 새로운 Language Representation Model
- 대형 코퍼스로 Unsupervised Learning을 통해 General Purpose Language Understanding Model을 구축하고, 이후 실행하고자 하는 task에 집중한 데이터로 Supervised Learning 하여 적용하는 Semi-Supervised Model
- 11개의 NLP Task에서 state-of-the-art를 기록하며 뉴욕 타임즈의 지면을 장식하기도 함

Nov. 18, 2018

Finally, a Machine That Can Finish Your Sentence

Completing someone else's thought is not an easy trick for A.I. But new systems are starting to crack the code of natural language.

By CADE METZ



BERT(Bidirectional Encoder Representations from Transformers)

- ELMo와 같이 양방향 문맥을 모두 고려하는 방식으로 접근
- 다만 ELMo는 Shallow Bidirectional 방식을 사용했으므로 보다 Deep한 접근법이 필요
- Masked Language Model(MLM)을 통해 Pre-Training하여 이러한 제약을 해결

```
Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon
```

- Transformer Model의 Encoder Part만을 사용하는 구조

BERT(Bidirectional Encoder Representations from Transformers)

- 다만 Pre-Training이 상당히 고가의 작업이므로 Kai.Lib Project에서 적용 가능할지는 의문
- SOTA급의 Word Embedding Method에 대한 이해 차원에서 논문 리뷰와 공부는 진행할 예정

> 사전학습(pre-training)은 상당히 고가로 4에서 16개의 Cloud TPU로 4일(12 층의 Transformer 모델의 경우 4개의 TPU를 사용하여 4일, 24층 Transformer 모델의 경우 16개의 TPU를 사용하여 4일이라는 의미) 각 언어마다 1회만의 순서이다. 자연 언어 처리 개발자는 처음부터 자신의 모델을 사전 학습할 필요가 없다.
>
> 전이학습(Fine-tuning)은 저렴하며, 논문(아래 참조)과 똑같은 사전학습이 끝난 모델을 사용하여 하나의 Cloud TPU를 이용, 1시간 GPU를 사용하면 2, 3시간만에 재현할 수 있다. 예를 들면 SQuAD는 하나의 Cloud TPU를 이용 30분으로 하나의 시스템으로서는 최첨단(state-of-the-art)인 91.0%의 Dev F1을 달성할 수 있다.

자료 출처 : 인공지능(AI) 언어모델 'BERT(버트)'는 무엇인가
- 인공지능신문(2019.01.03)

참고할 만 한 책