딥러닝 자연어처리

RNN에서 BERT까지

우 종 하

발표자 소개

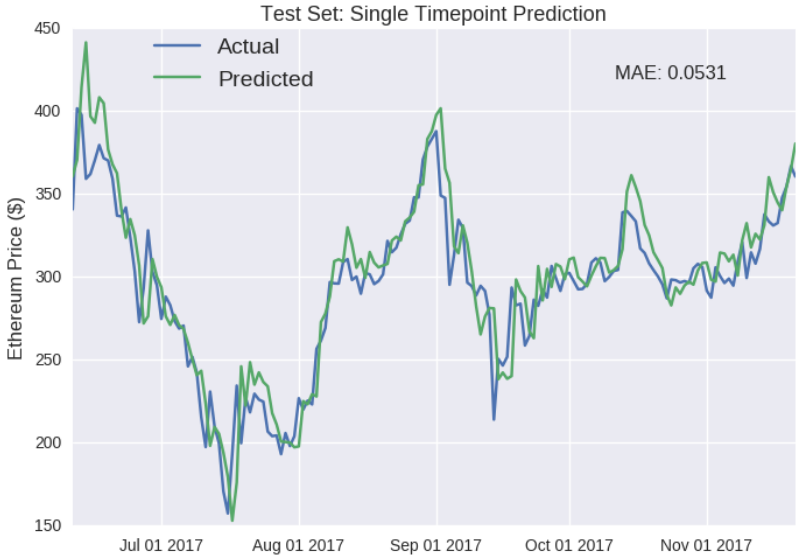
- 챗봇 개발자 모임 운영
  - 페이스북 그룹
  - 챗봇, 인공지능 스피커, 자연어 처리 관련 커뮤니티

https://www.facebook.com/groups/ChatbotDevKR

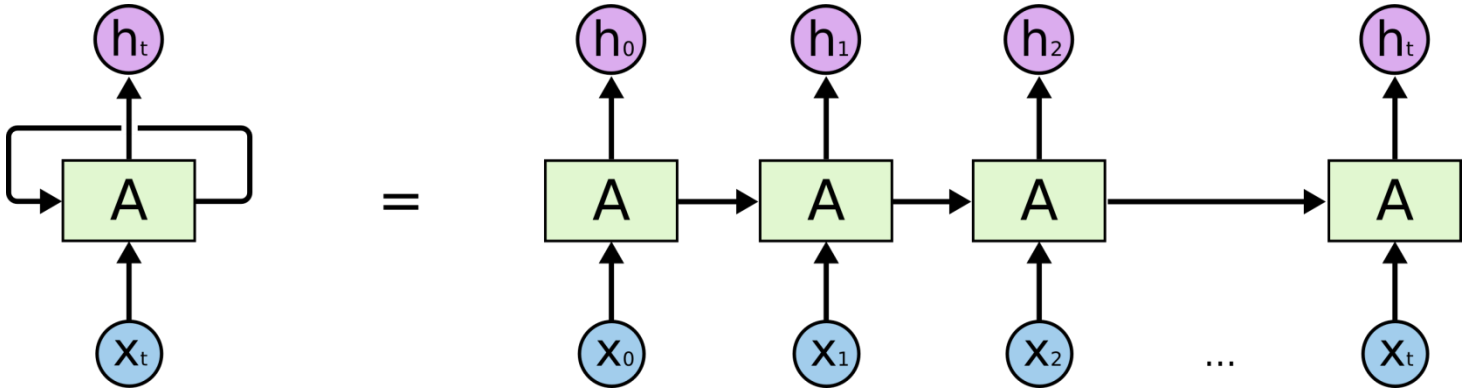
RNN
(Recurrent Neural Network)

기존 신경망의 한계

- 연속적인 시퀀스를 처리하기 어려움
- 입력의 순서가 중요한 분야
  - 자연어처리
  - 음성인식
  - 주식
  - 날씨
  - 음악

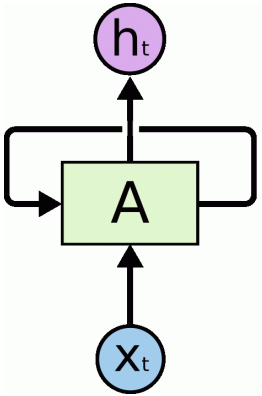


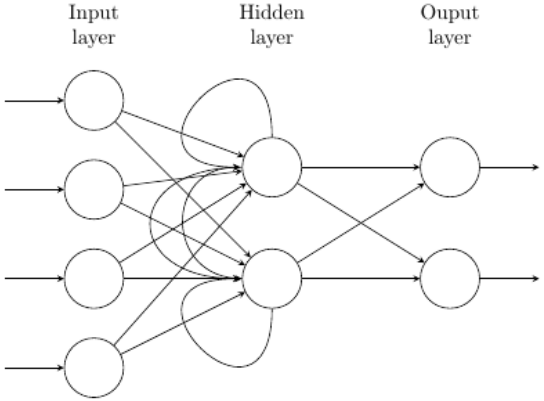
RNN 구조



이전 출력값이 현재 결과에 영향을 미침

RNN 구조

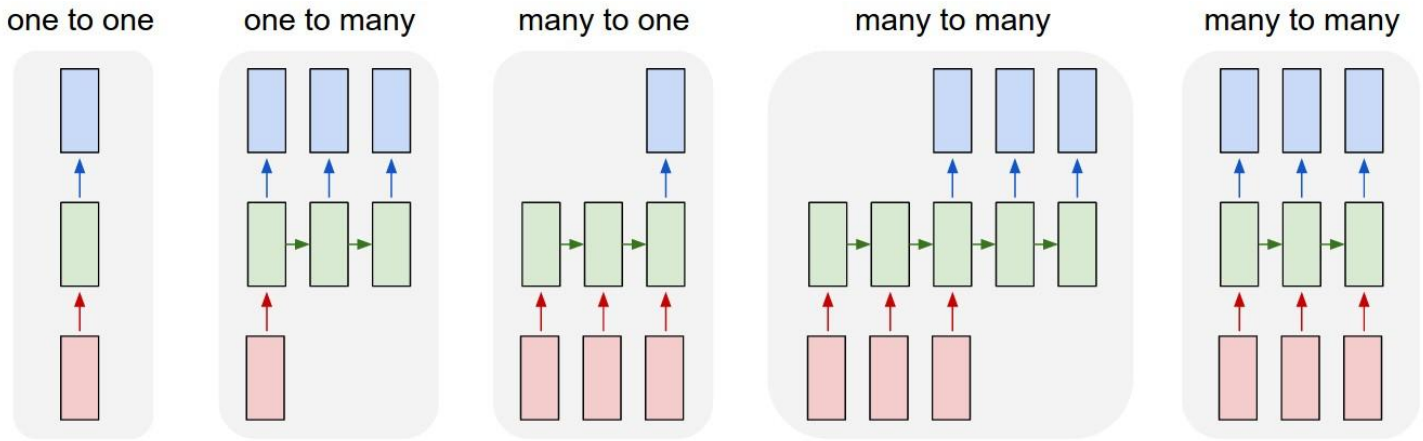




$$
\begin{aligned}
\boldsymbol{a}^{(t)} &= \boldsymbol{b} + \boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{U}\boldsymbol{x}^{(t)} \\
\boldsymbol{h}^{(t)} &= \tanh(\boldsymbol{a}^{(t)}) \\
\boldsymbol{o}^{(t)} &= \boldsymbol{c} + \boldsymbol{V}\boldsymbol{h}^{(t)} \\
\hat{\boldsymbol{y}}^{(t)} &= \mathrm{softmax}(\boldsymbol{o}^{(t)})
\end{aligned}
$$

순환 W와 입력 U의 두 개의 가중치가 존재

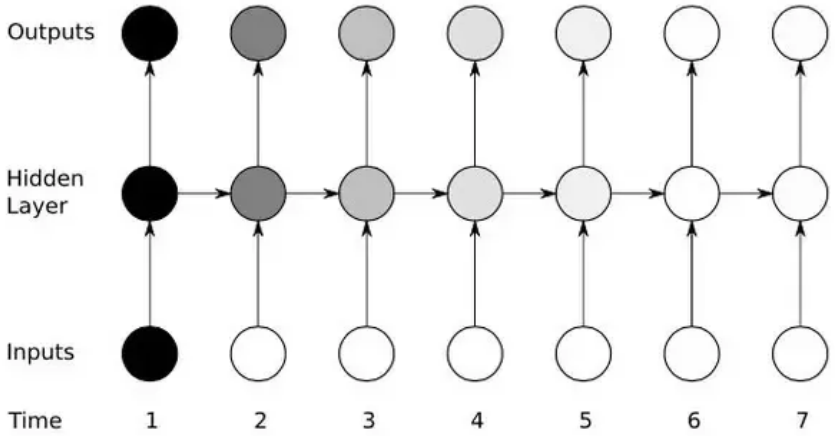
RNN 구조



입력과 출력에서 다양한 형태 가능

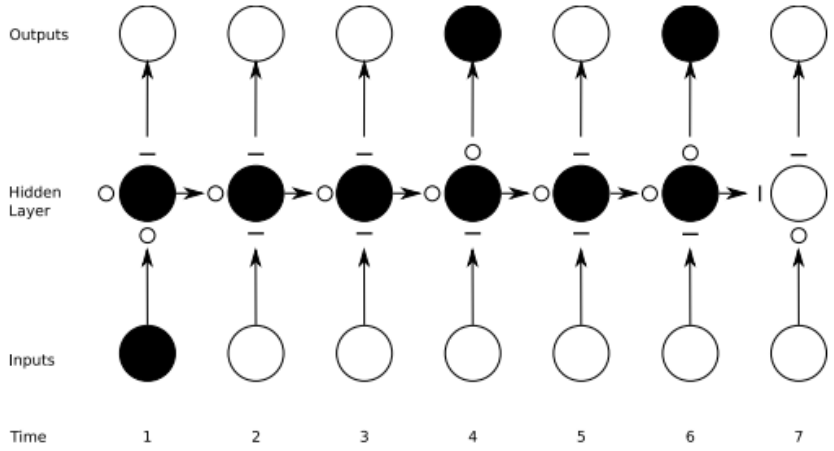
LSTM
(Long Short Term Memory)

RNN의 문제점



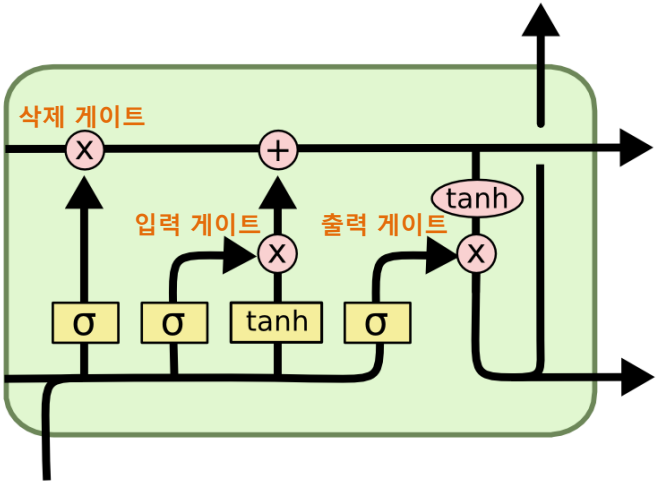
처음 입력의 정보가 뒤로 갈수록 사라짐

LSTM의 목적



입력 중 핵심적인 정보를 잊어버리지 않고 뒤로 전달

LSTM 구조



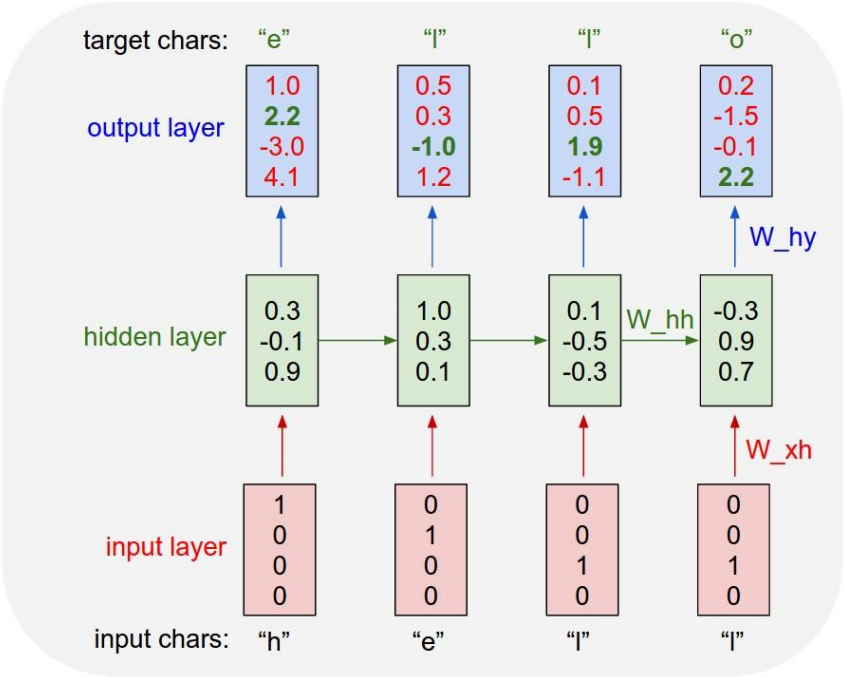
$$
\begin{aligned}
\mathbf{i}_{(t)} &= \sigma(\mathbf{W}_{xi}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hi}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{i}) \\
\mathbf{f}_{(t)} &= \sigma(\mathbf{W}_{xf}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hf}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{f}) \\
\mathbf{o}_{(t)} &= \sigma(\mathbf{W}_{xo}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{ho}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{o}) \\
\mathbf{g}_{(t)} &= \tanh(\mathbf{W}_{xg}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hg}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{g}) \\
\mathbf{c}_{(t)} &= \mathbf{f}_{(t)} \otimes \mathbf{c}_{(t-1)} + \mathbf{i}_{(t)} \otimes \mathbf{g}_{(t)} \\
\mathbf{y}_{(t)} &= \mathbf{h}_{(t)} = \mathbf{o}_{(t)} \otimes \tanh(\mathbf{c}_{(t)})
\end{aligned}
$$

입력과 순환 각각 4개씩 총 8개의 가중치가 존재

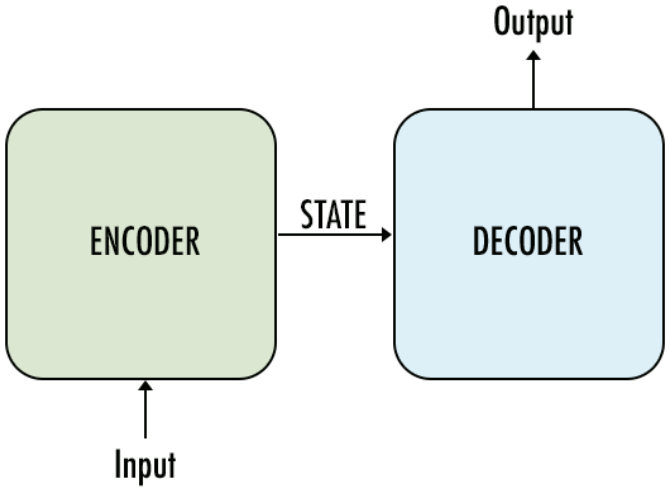
Seq2Seq 모델

RNN 문장 생성의 문제점

- 출력이 바로 이전 입력까지만 고려해서 정확도 떨어짐
- 전체 입력 문장을 반영하지 못함

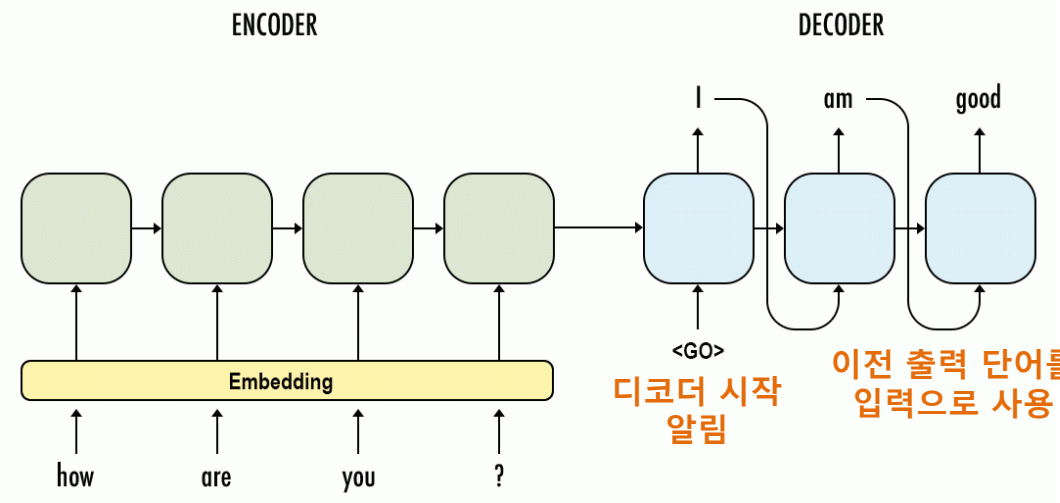


Seq2Seq 모델 구조



인코더와 디코더 두 개의 LSTM으로 구성

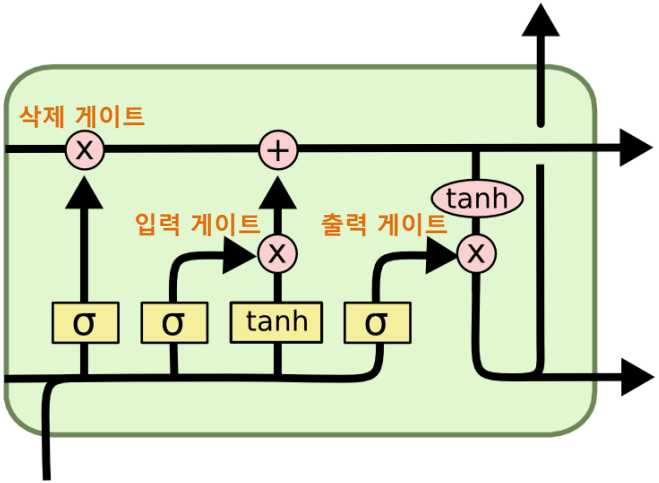
Seq2Seq 모델 구조



인코더로 입력 문장을 먼저 처리하고 디코더로 답변 문장 출력

어텐션
(Attention)

LSTM Seq2Seq 모델의 한계

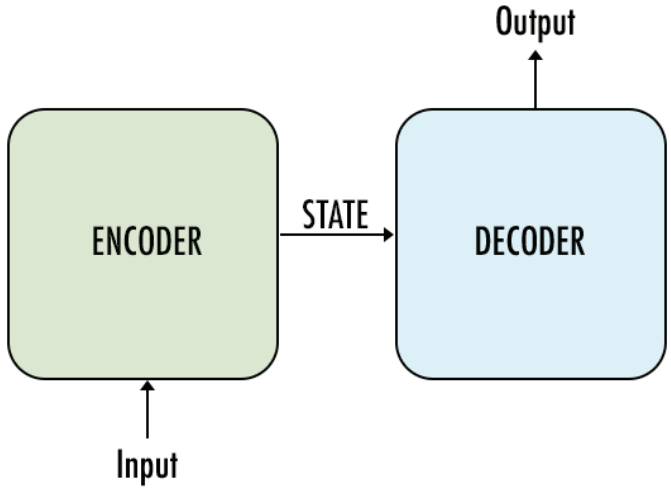


$$
\begin{aligned}
\mathbf{i}_{(t)} &= \sigma(\mathbf{W}_{xi}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hi}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{i}) \\
\mathbf{f}_{(t)} &= \sigma(\mathbf{W}_{xf}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hf}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{f}) \\
\mathbf{o}_{(t)} &= \sigma(\mathbf{W}_{xo}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{ho}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{o}) \\
\mathbf{g}_{(t)} &= \tanh(\mathbf{W}_{xg}{}^{T} \cdot \mathbf{x}_{(t)} + \mathbf{W}_{hg}{}^{T} \cdot \mathbf{h}_{(t-1)} + \mathbf{b}_{g}) \\
\mathbf{c}_{(t)} &= \mathbf{f}_{(t)} \otimes \mathbf{c}_{(t-1)} + \mathbf{i}_{(t)} \otimes \mathbf{g}_{(t)} \\
\mathbf{y}_{(t)} &= \mathbf{h}_{(t)} = \mathbf{o}_{(t)} \otimes \tanh(\mathbf{c}_{(t)})
\end{aligned}
$$

정보의 흐름을 조정하는 게이트만으로는 부족

LSTM Seq2Seq 모델의 한계

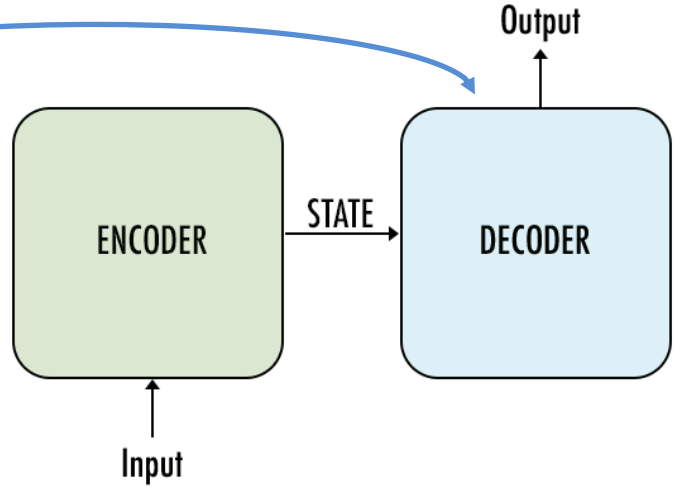
강다니엘은 워너원의 센터인데 키도 크고 멋지고 잘생겼어. 너무 사랑스럽다. 노래도 잘하고 춤도 잘추고 완벽한 남자야. 다시 한번 말하지만 워너원의 센터는 누구?



입력 문장이 길어지면 답변의 정확도 떨어짐

어텐션의 원리

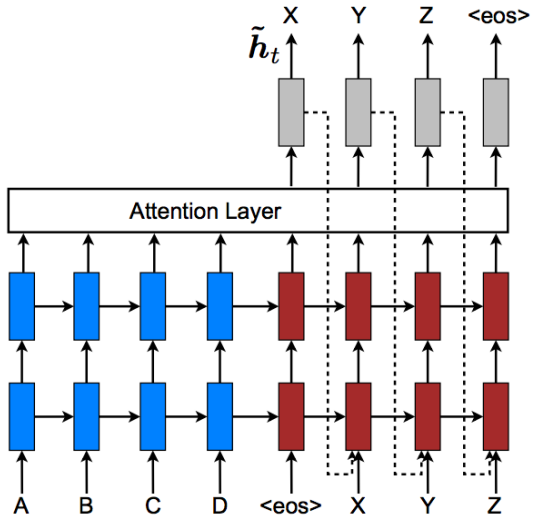
강다니엘은 워너원의 센터인데 키도 크고 멋지고 잘생겼어. 너무 사랑스럽다. 노래도 잘하고 춤도 잘추고 완벽한 남자야. 다시 한번 말하지만 워너원의 센터는 누구?



중요한 단어에 집중(attention)하여 디코더에 바로 전달

어텐션 Seq2Seq 모델 구조

디코더에서 각 단어를 생성할 때
인코더의 어떤 단어에서 정보를 받을지
어텐션 레이어가 결정



인코더의 출력값들을 모아 디코더 계산에 같이 사용

트랜스포머
(Transformer)

어텐션만으로 충분하다

- 2017년 구글이 발표
- LSTM 필요 없음
- 어텐션 신경망만 사용하여 인코더-디코더 구현

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

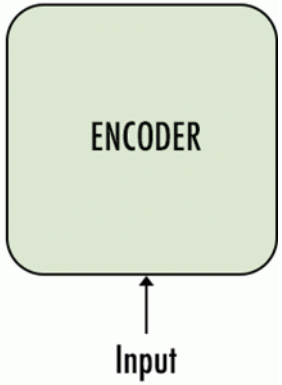
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

트랜스포머의 원리

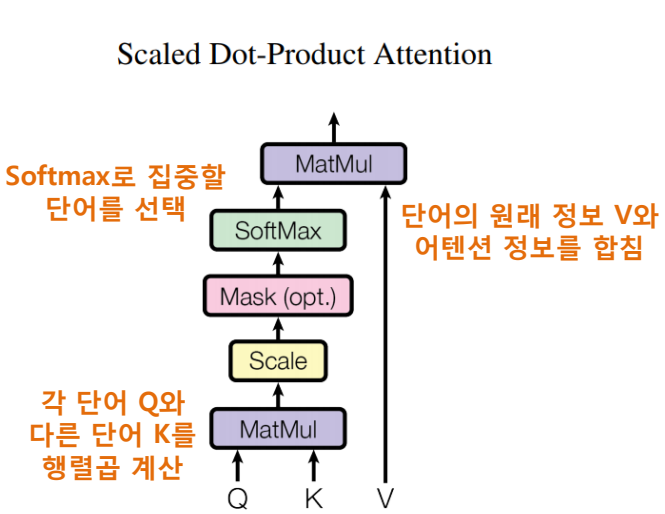
강다니엘은 워너원의 센터인 데 키도 크고 멋지고 잘생겼어. 너무 사랑스럽다. 노래도 잘하고 춤도 잘추고 완벽한 남자야. 다시 한번 말하지만 워너원의 센터는 누구?

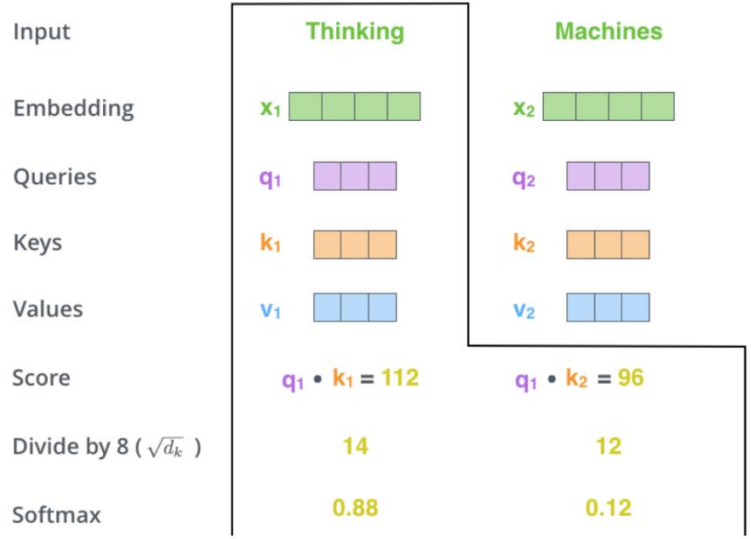
강다니엘은 워너원의 센터인 데 키도 크고 멋지고 잘생겼어. 너무 사랑스럽다. 노래도 잘하고 춤도 잘추고 완벽한 남자야. 다시 한번 말하지만 워너원의 센터는 누구?



인코더에서 디코더가 아니라 스스로 셀프 어텐션

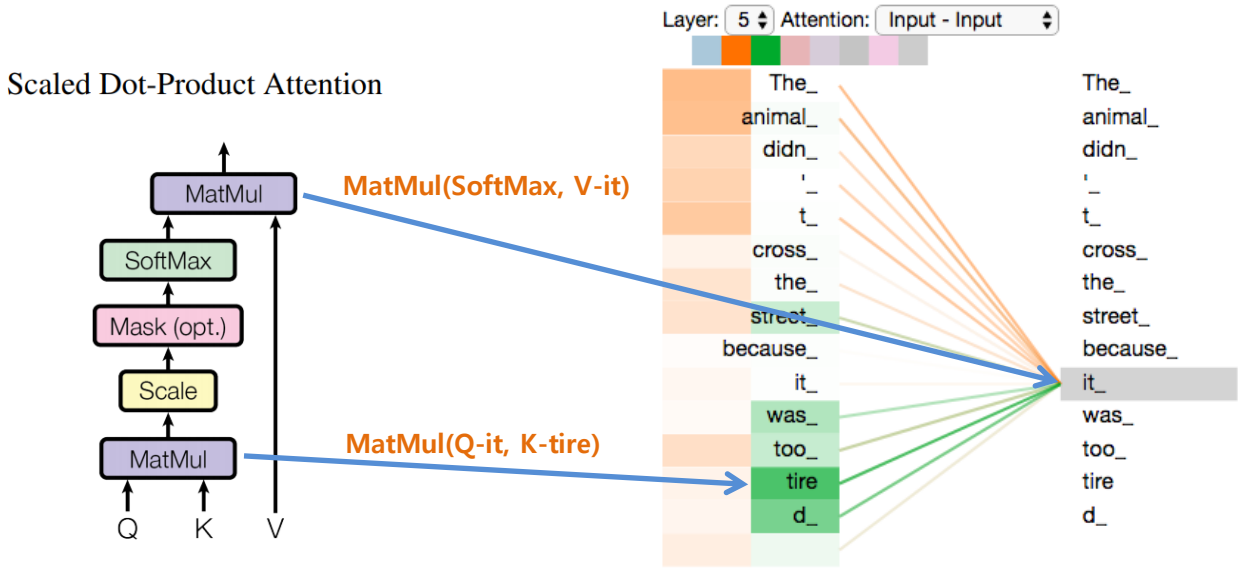
셀프 어텐션





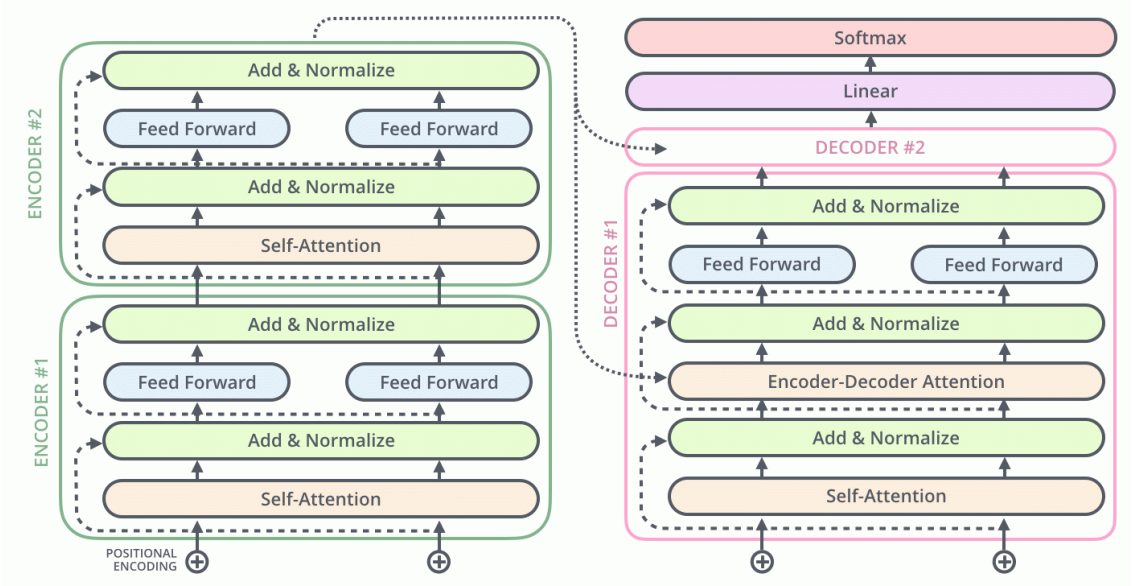
각 단어를 Q K V로 변환하여 어텐션 계산

셀프 어텐션



문장에서 중요한 단어들에 집중하여 각 단어의 정보 업데이트

트랜스포머 구조



인코더와 디코더는 각각 여러 개로 중첩되어 구성

BERT
(Bidirectional Encoder Representations from Transformers)

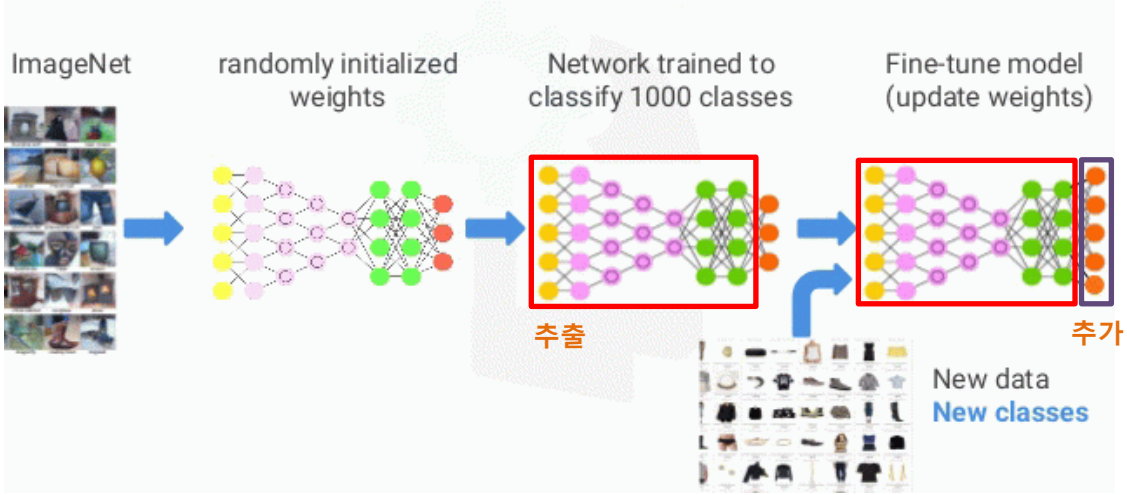
BERT 소개

- 2018년 10월 구글이 발표
- 사전 훈련 기반 딥러닝 언어 모델
- 트랜스포머로 구현됨
- 다양한 자연어처리 분야에 응용 가능

사전 훈련(Pre-training)

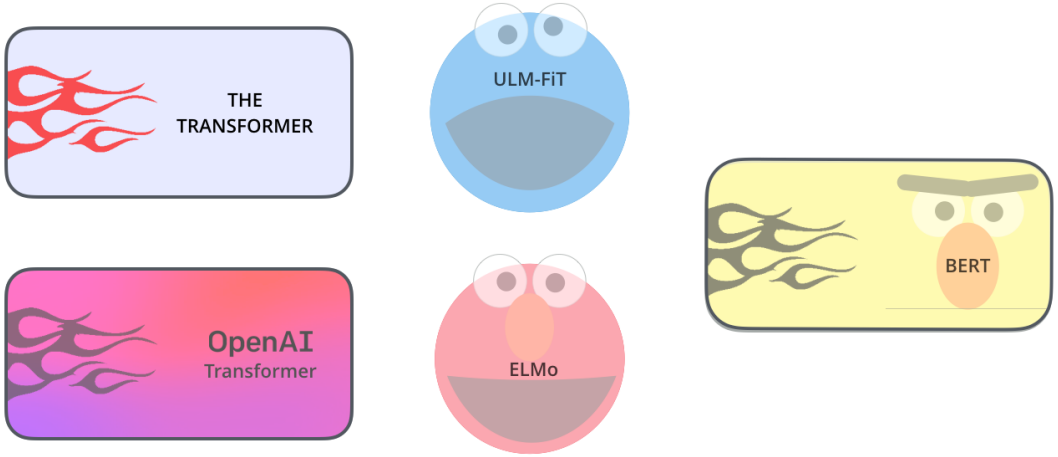
- 이미지 학습에 많이 사용됨
- 방법
  - 사진과 라벨이 있는 대용량의 데이터를 학습한 모델 생성
  - 모델에서 입력과 가까운 부분은 일반적인 패턴 감지 가능
  - 이 부분을 추출하고 출력 부분에 신경망을 추가하여 새로운 모델 생성
  - 자신만의 데이터로 학습 수행
- 장점
  - 적은 데이터로 더 빨리 학습 가능

전이 학습(Transfer Learning)



사전 훈련 모델로 새로운 모델을 만들어 다시 학습

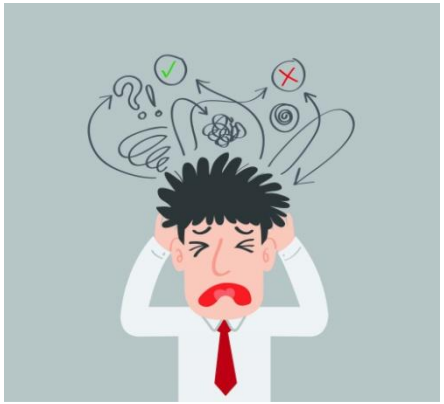
자연어처리에 사전 훈련 적용



2018년에 사전 훈련 기반의 언어 모델들이 발표됨

사전 훈련의 필요성

하이 -> 헬로우
나이스투미튜 -> 유튜
땡큐 -> 유어웰컴
아임쏘리 -> 아임파인
웨어아유프롬 -> 아임프롬서울
...



처음부터 무식하게 외우는 것은 어렵고 오래 걸림

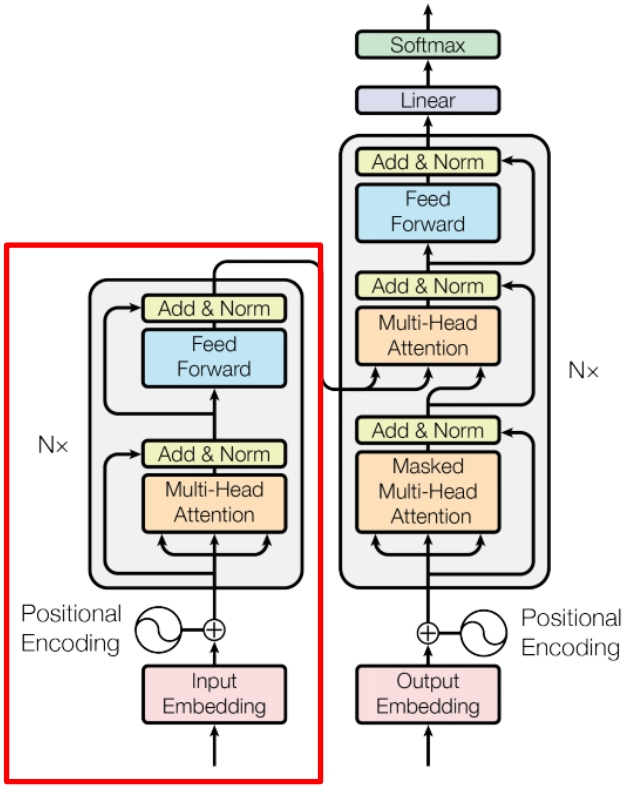
사전 훈련의 필요성

Hi -> Hello
Nice to meet you -> You too
Thank you -> You're welcome
I'm sorry -> I'm fine
Where are you from -> I'm from Seoul
...

먼저 언어의 기본을 익히고 문장을 외우면 더 빠르고 효과적

BERT 구조

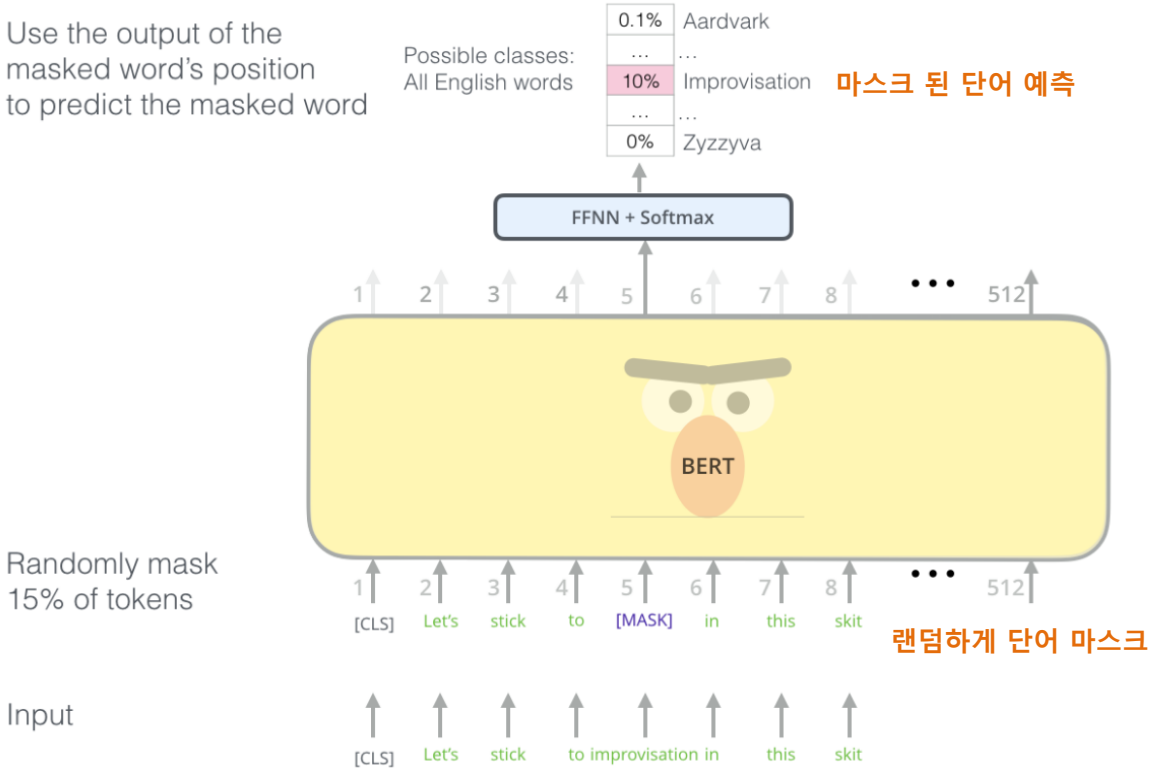
- 트랜스포머의 인코더만 사용
- BERT Base
  - 12개의 트랜스포머 블록
- BERT Large
  - 24개의 트랜스포머 블록



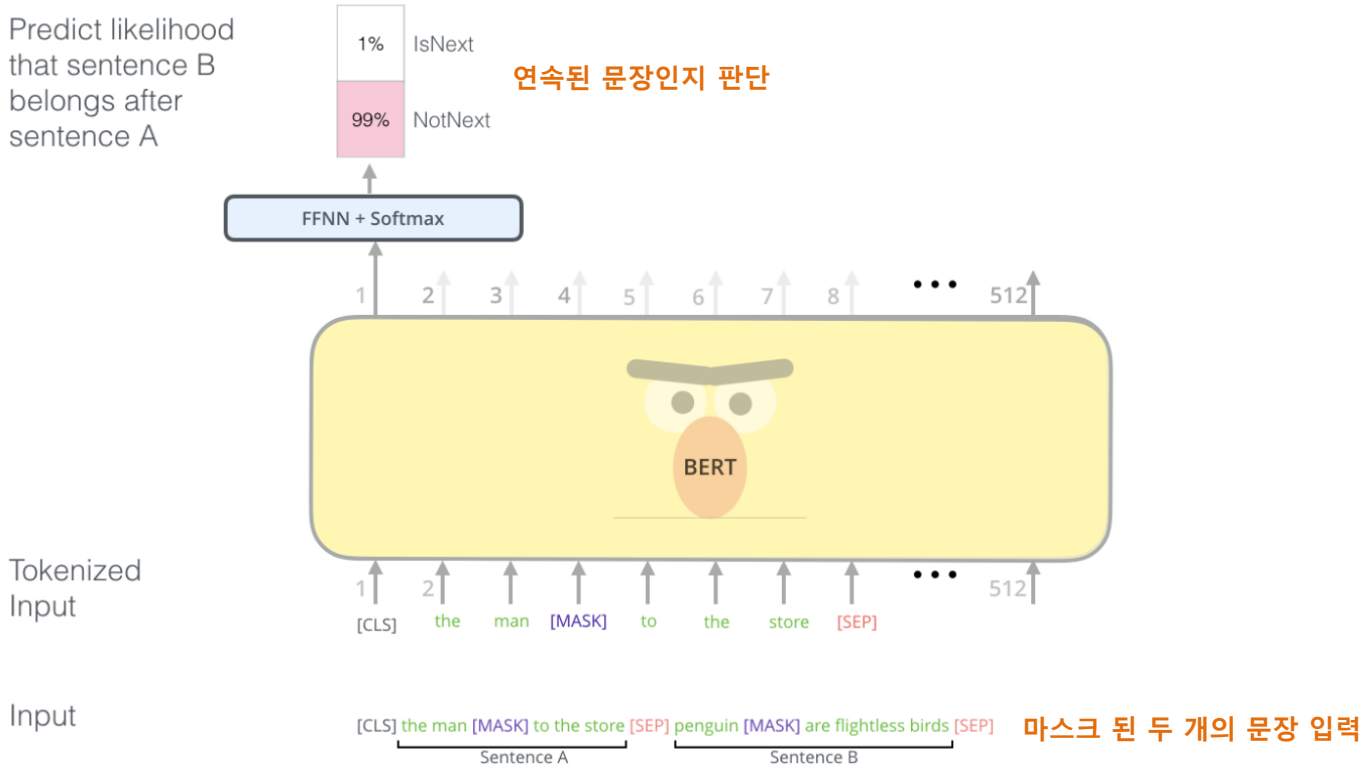
사전 훈련 데이터

- BooksCorpus
  - 8억 단어
- Wikipedia
  - 25억 단어

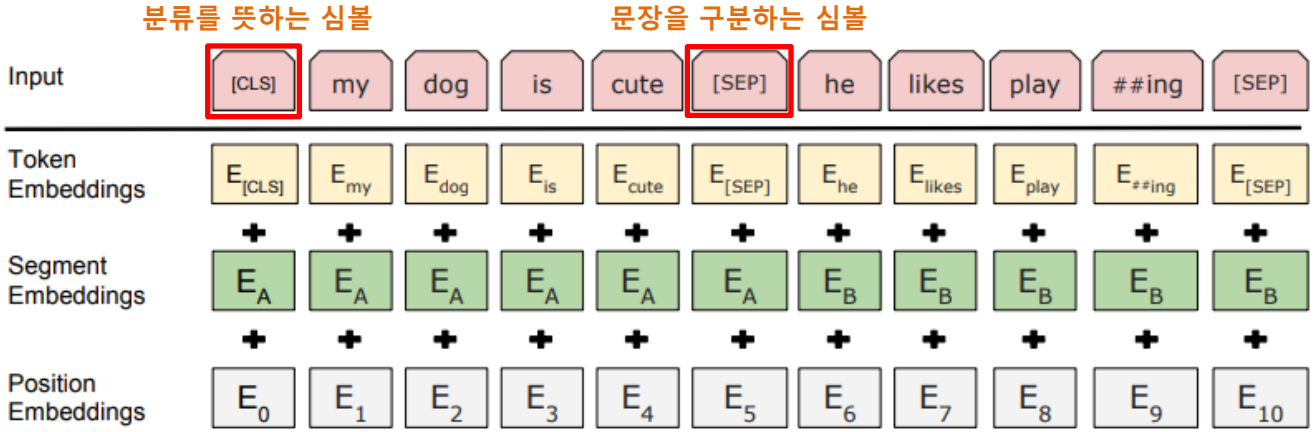
사전 훈련 방법 1 - 단어 마스크



사전 훈련 방법 2 - 다음 문장 예측



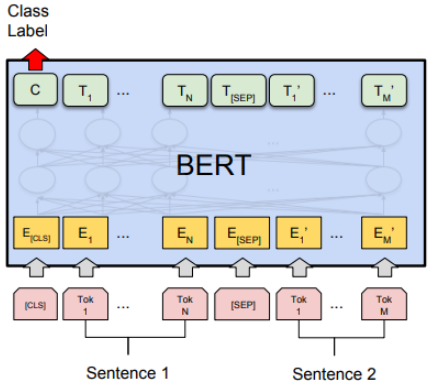
입력 임베딩

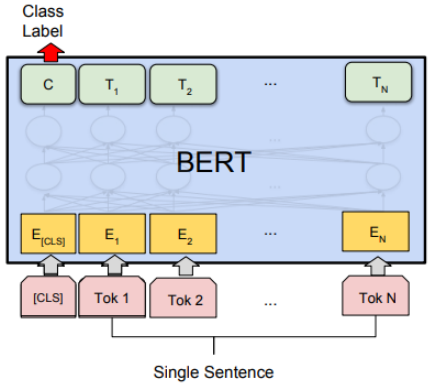


3개 임베딩의 총합을 입력으로 사용

BERT 모델로 전이 학습

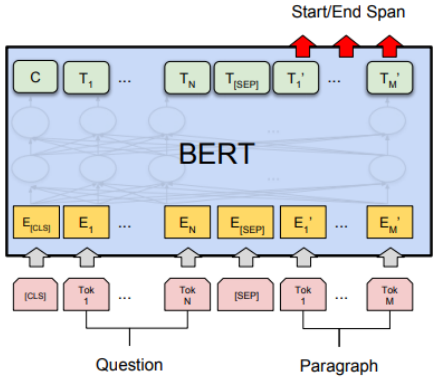
두 문장 분류

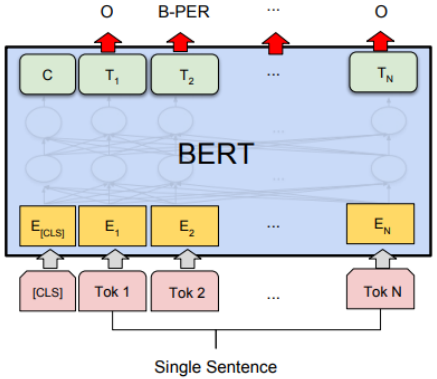




한 문장 분류

질의 응답





문장 태깅

SQuAD 질의 응답 대회

The Stanford Question Answering Dataset

Steam engines are external combustion engines, where the working fluid is separate from the combustion products. Non-combustion heat sources such as solar power, nuclear power or geothermal energy may be used. The ideal thermodynamic cycle used to analyze this process is called the Rankine cycle. In the cycle, water is heated and transforms into steam within a boiler operating at a high pressure. When expanded through pistons or turbines, mechanical work is done. The reduced-pressure steam is then condensed and pumped back into the boiler.

Along with geothermal and nuclear, what is a notable non-combustion heat source?
Ground Truth Answers: solar | solar power | solar power, nuclear power or geothermal energy | solar

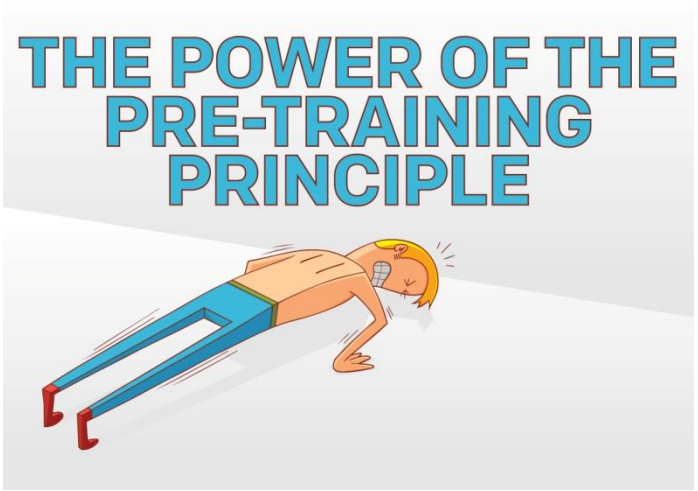
질문과 지문이 주어지면 정답을 찾음

SQuAD 질의 응답 대회

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	**83.536**	**86.096**
2 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
3 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035
4 Dec 16, 2018	PAML+BERT (single model) *PINGAN GammaLab*	82.577	85.603
5 Nov 16, 2018	AoA + DA + BERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.374	85.310

BERT가 발표된 후 바로 상위권 독차지

딥러닝 자연어처리의 발전 방향



사전 훈련 모델이 더욱 중요해질 가능성 높음

감사합니다