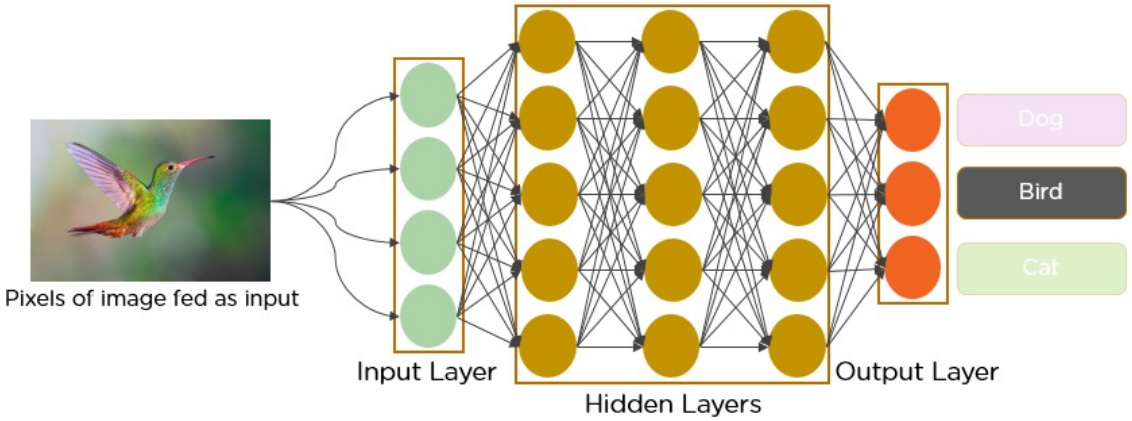
딥러닝과 자연어 처리:
어텐션 메커니즘과 트랜스포머 역할

Table of Contents
목차

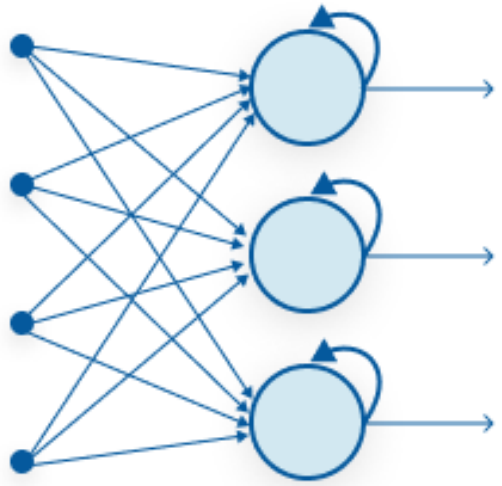
딥러닝이란?

딥 러닝 (Deep Learning)은 머신러닝 (Machine Learning)의 특정한 한 분야로서 인공 신경망(Artificial Neural Network)의 층을 연속적으로 깊게 쌓아올려 데이터를 학습하는 방식.

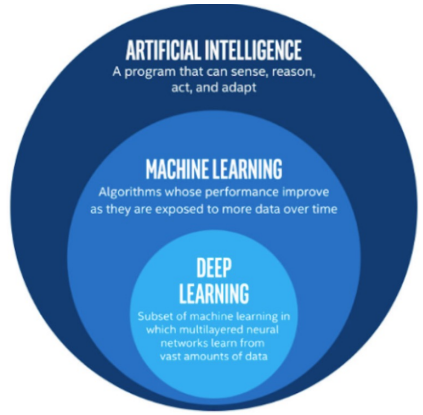
- 다층 퍼셉트론, CNN, RNN, 등 다양한 아키텍쳐가 존재합니다.



Convolutional Neural Network



Recurrent Neural Network



딥러닝 역사

"The foundations for all of this artificial intelligence were laid at Cornell."

– Thorsten Joachims, professor in CIS

1세대: 1958 "Perceptron" [코넬 항공 연구소 프랭크 로젠블랫]

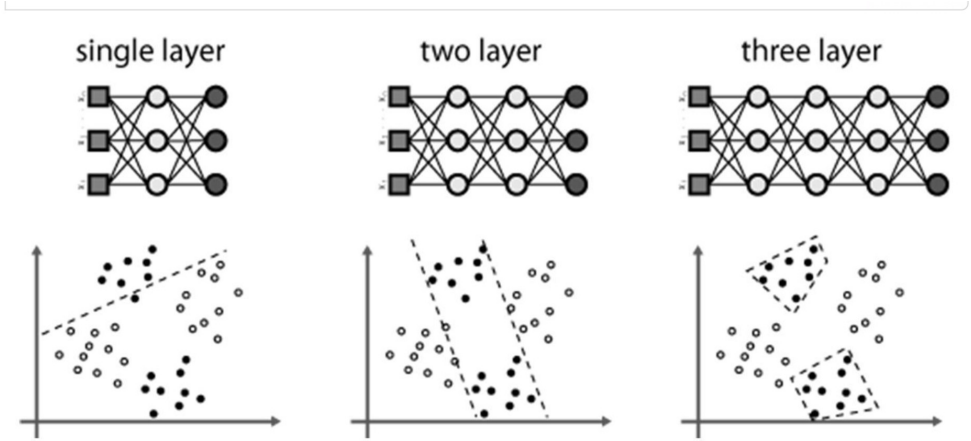
- 뇌의 뉴런이 서로 연결되어 정보를 처리하는 구조를 모방한 알고리즘

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b)$$

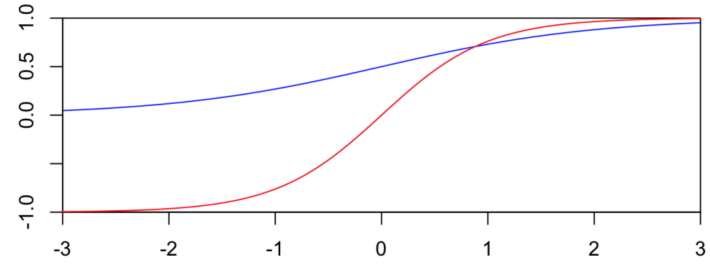
2세대: Multilayer Perceptron

- Hidden layer추가

3세대: Supervised Learning – Rectified linear unit (ReLU), Dropout



Multilayer Perceptron

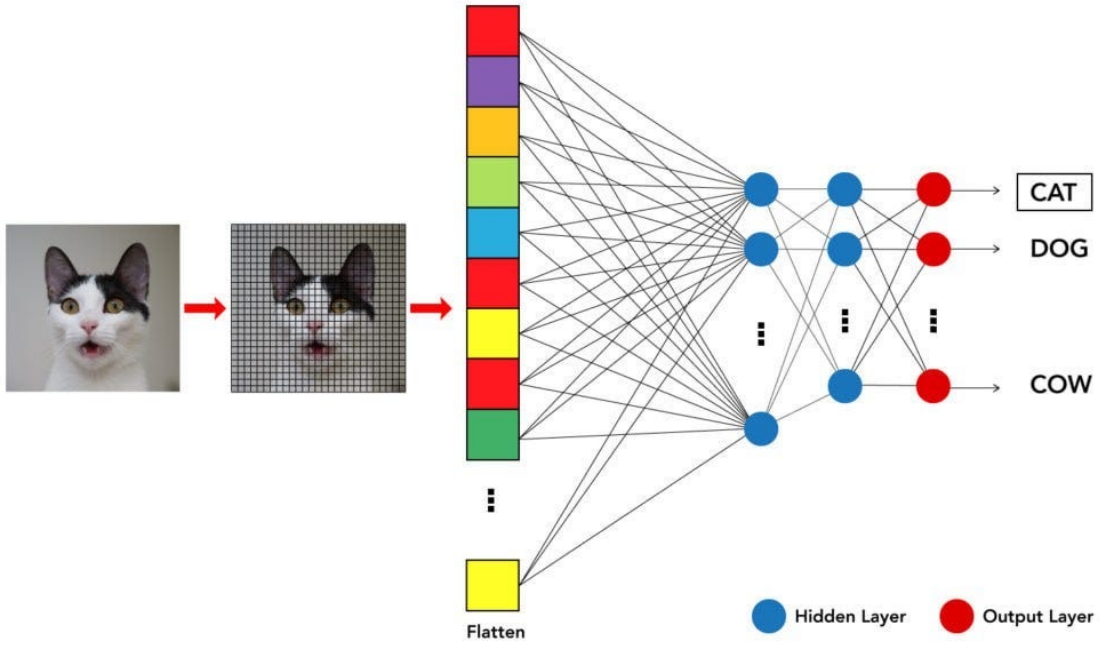


Sigmoid functions

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

입력 계층 / 은닉 계층 / 출력 계층

- 주요 응용 분야로는 이미지 인식, 자연어 처리, 자율 주행 등이 있습니다.
- 딥러닝의 핵심은 다층 구조로 복잡한 패턴을 학습화 할 수 있다는 점입니다.
- 예를 들어, 이미지 인식에서는 여러 층을 통해 점차 복잡한 특징을 추출합니다.

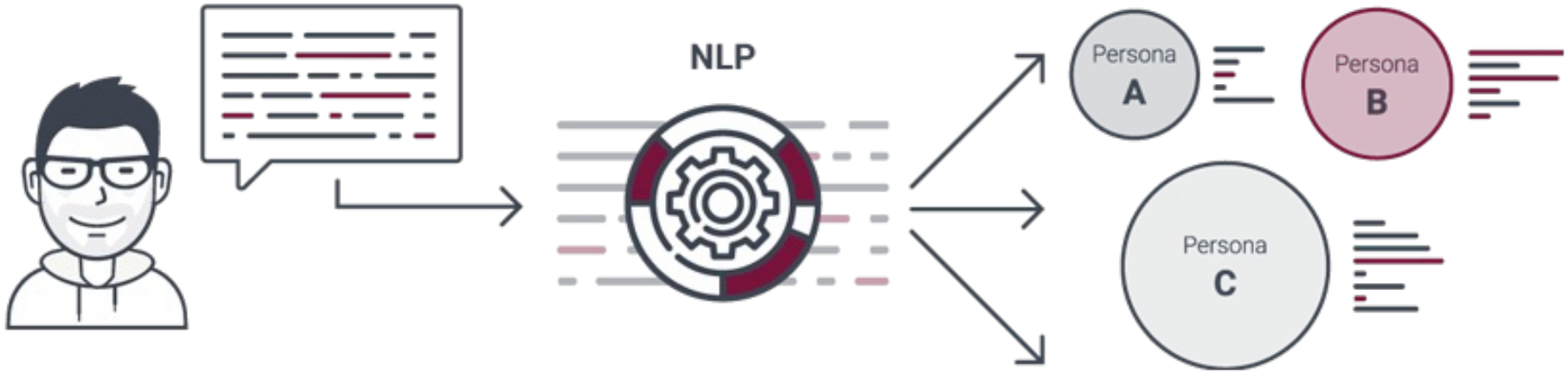


Natural Language Processing (NLP)란 무엇인가요?

자연어 처리(NLP)는 인간의 언어를 해석, 조작 및 이해하는 능력을 컴퓨터에 부여하는 AI 기술입니다.

NLP 성능:
- 텍스트 및 음성 데이터를 완벽하고 효율적으로 분석하는데 중요
- 언어 속 방언, 속어, 문법적 불규칙성의 차이를 극복할 수 있다.

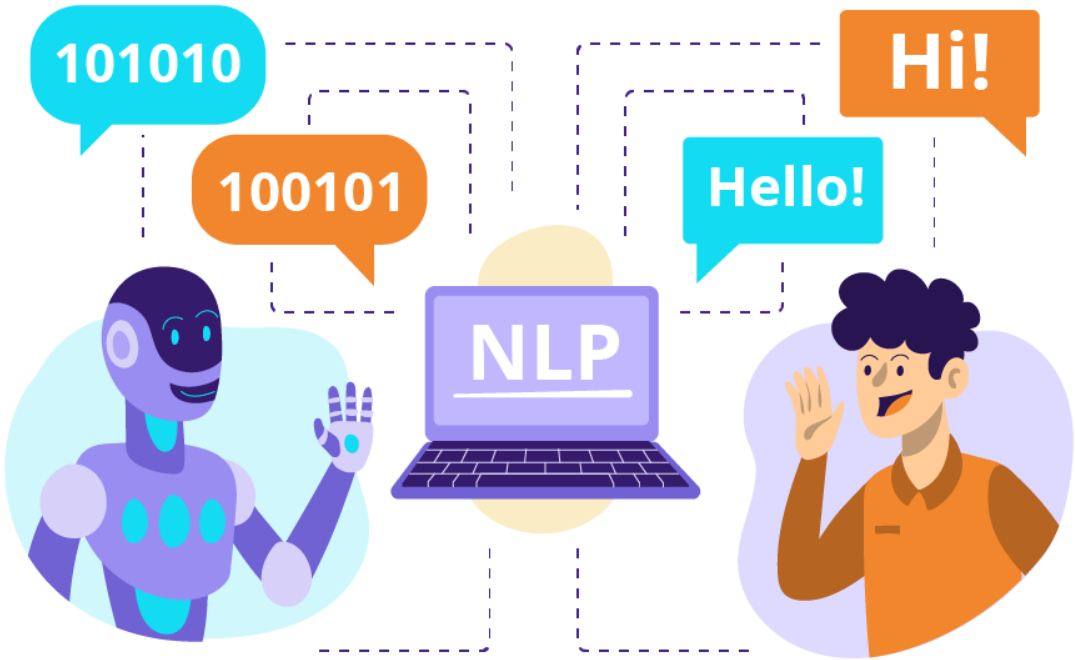
작업에서는 구글 번역과 같은 서비스가 NLP를 사용하여 다양한 언어를 지원합니다.

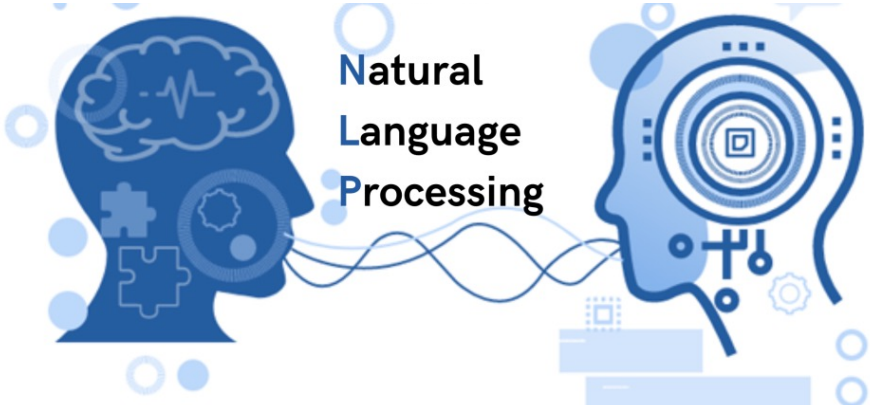


NLP의 발전

- 최종목표: 인간과 AI의 상호작용을 자연스럽고 의미 있게 만드는 것
- 주요 도전과제: 방대한 데이터셋 처리, 언어의 모호성 이해, 언어의 문맥 이해, 확장 가능한 모델 구축
  - 예) “나는 은행에서 물고기를 잡았다”

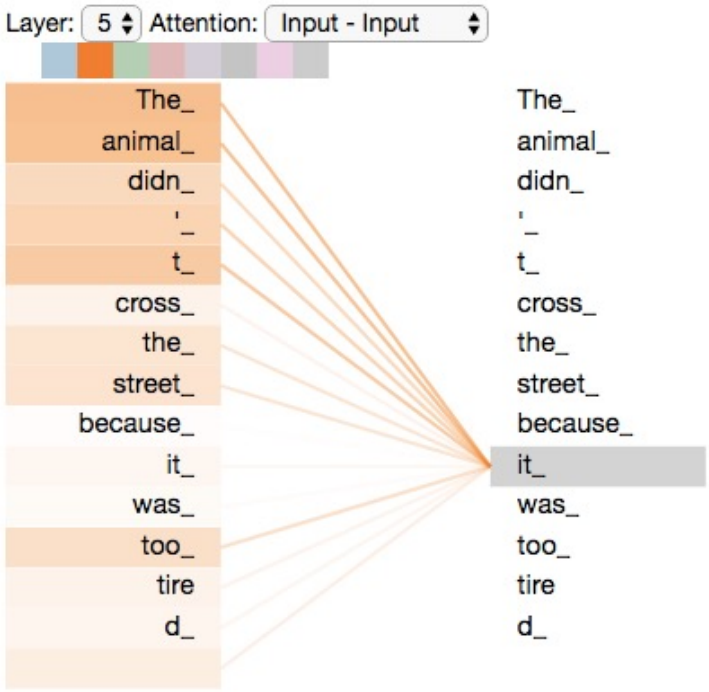
- 이러한 문제를 해결하기 위해 어텐션 메커니즘이 도입되었습니다.





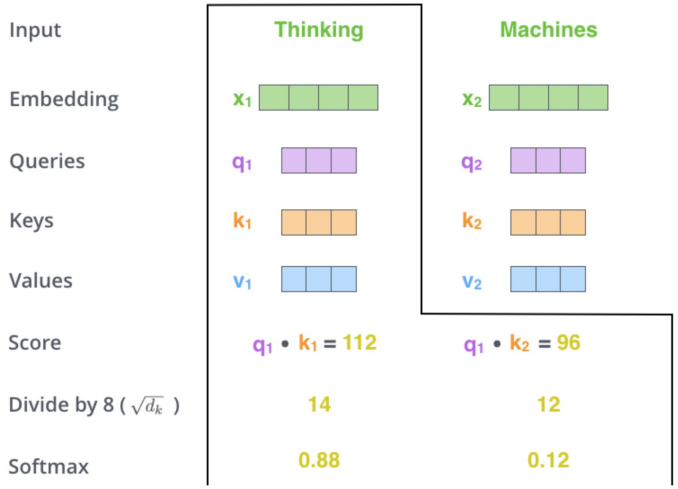
어텐션 (Attention) 소개 / 필요성

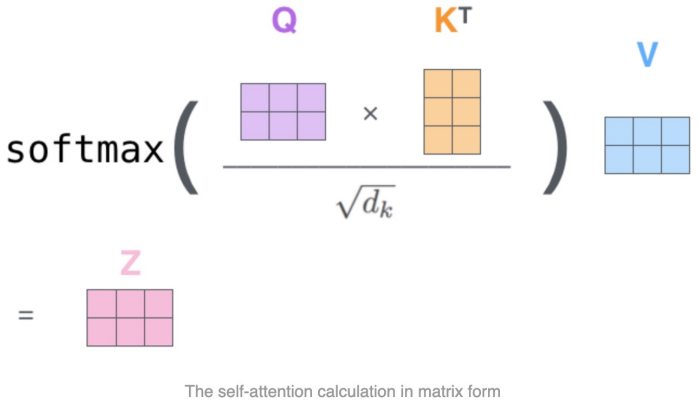
- 어텐션 메커니즘은 입력 시퀀스의 관련 부분에 집중할 수 있게 하는 기술입니다.
- 이 기술은 모델이 중요한 정보에 더 많은 가중치를 부여할 수 있도록 도와줍니다.
- 자연어 처리에서 어텐션은 문맥을 이해하는데 큰 도움을 줍니다.



어텐션 메커니즘 작동 원리

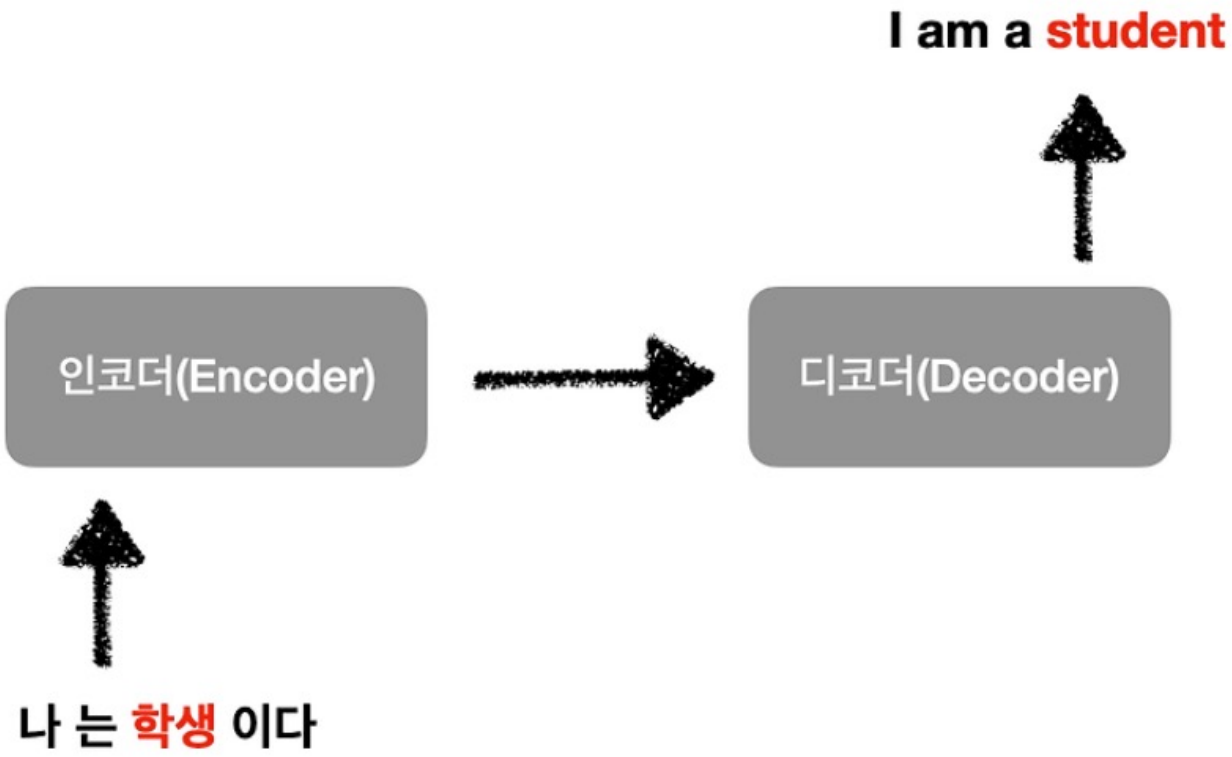
- Query, Key, Value:
  - Query: 현재 주목하고 있는 부분
  - Key: 모든 입력 부분
  - Value: Key와 연관된 실제 값
- Attention 점수 계산:
  - Query 와 Key의 유사도를 계산하여 얻음
- 가중치 계산:
  - 소프트맥스 함수를 통해 가중치를 계산
- 가중합 계산:
  - 가중치를 통해 Value의 가중합을 구하여 중요한 정보에 집중



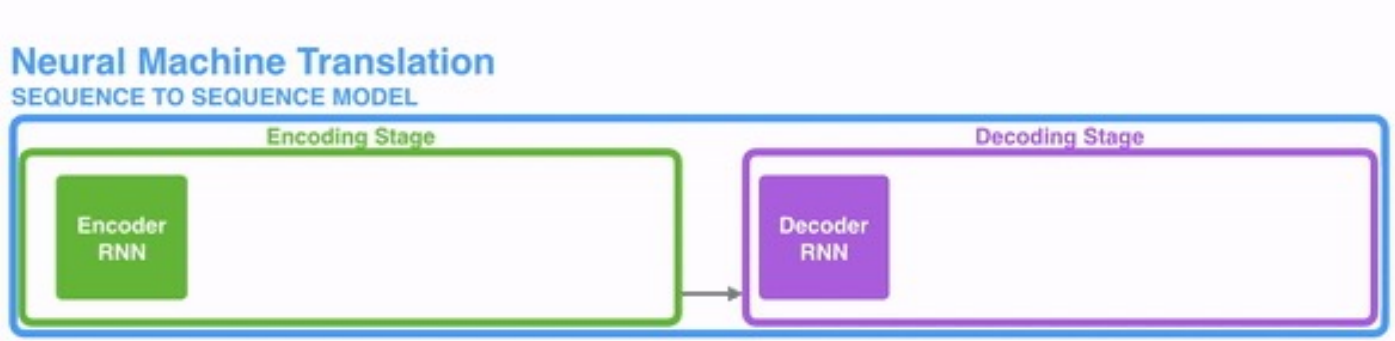


The self-attention calculation in matrix form

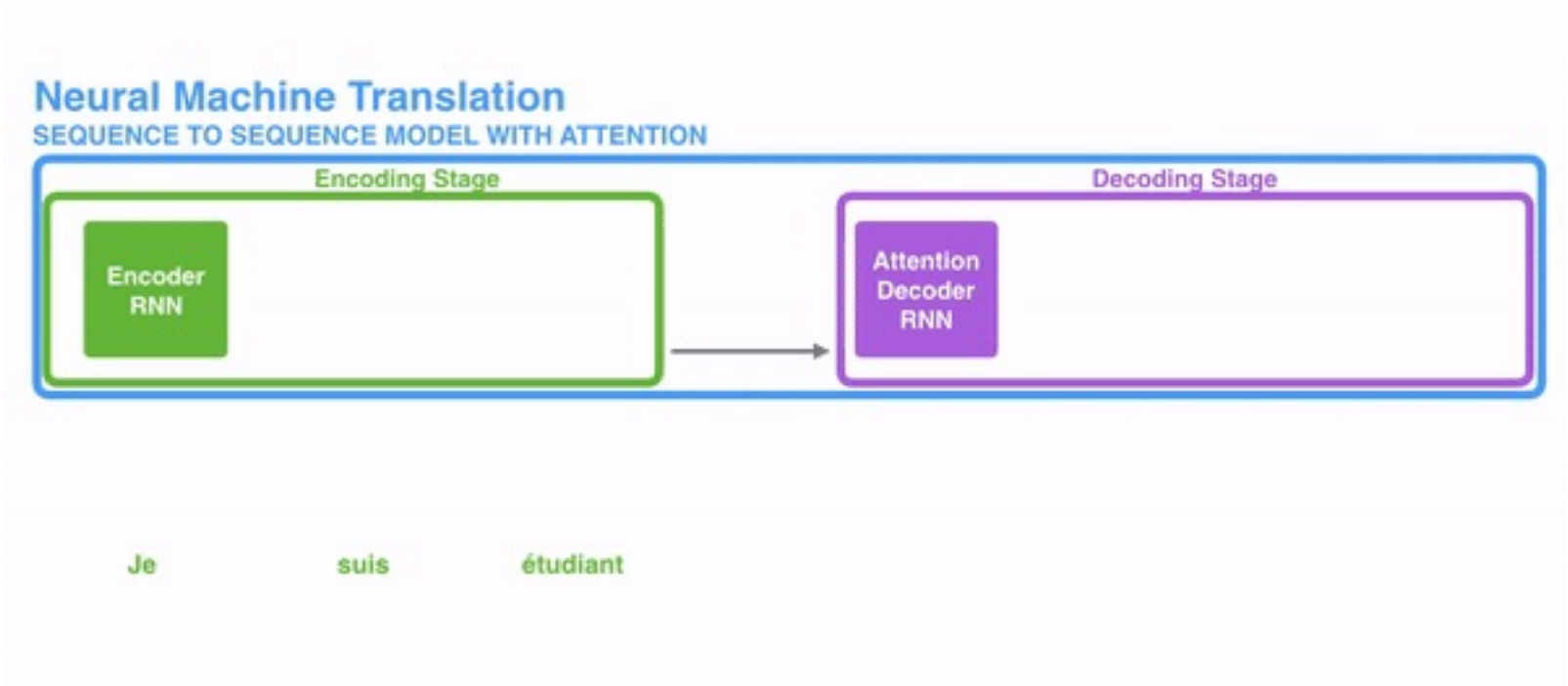
어텐션 메커니즘 작동 예시



어텐션 메커니즘

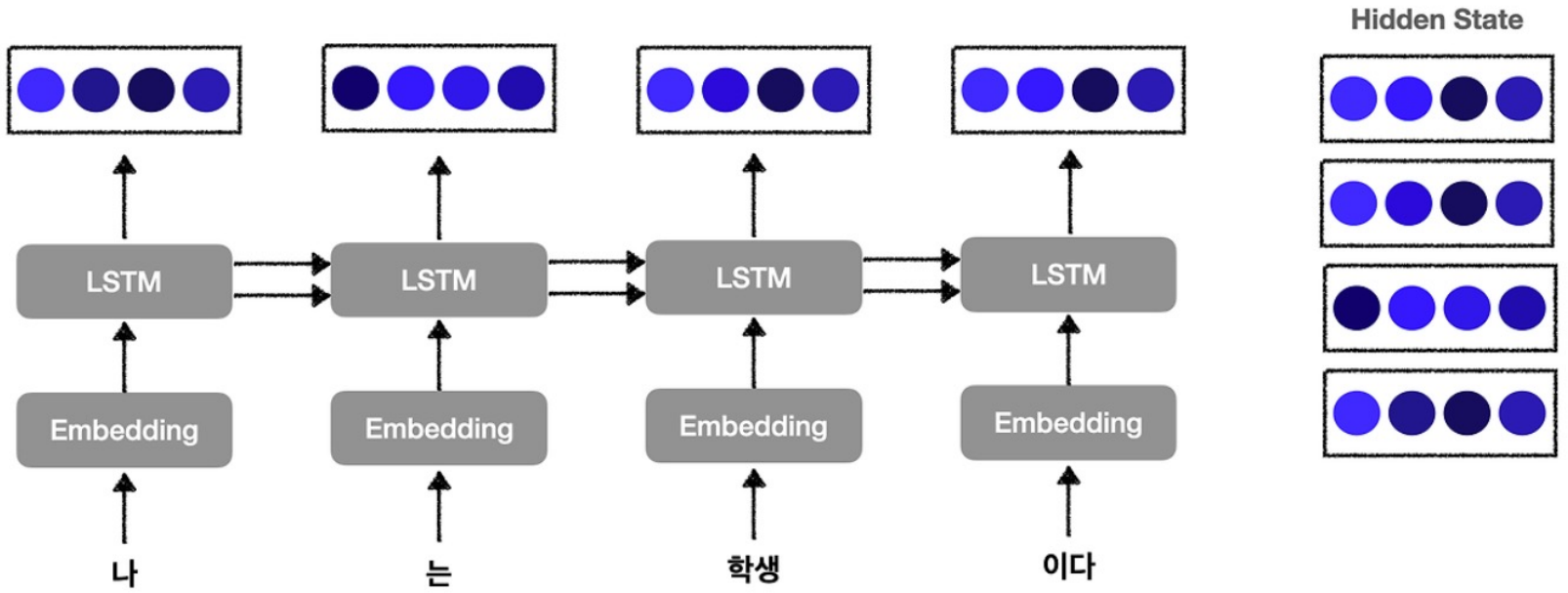


Seq2seq model

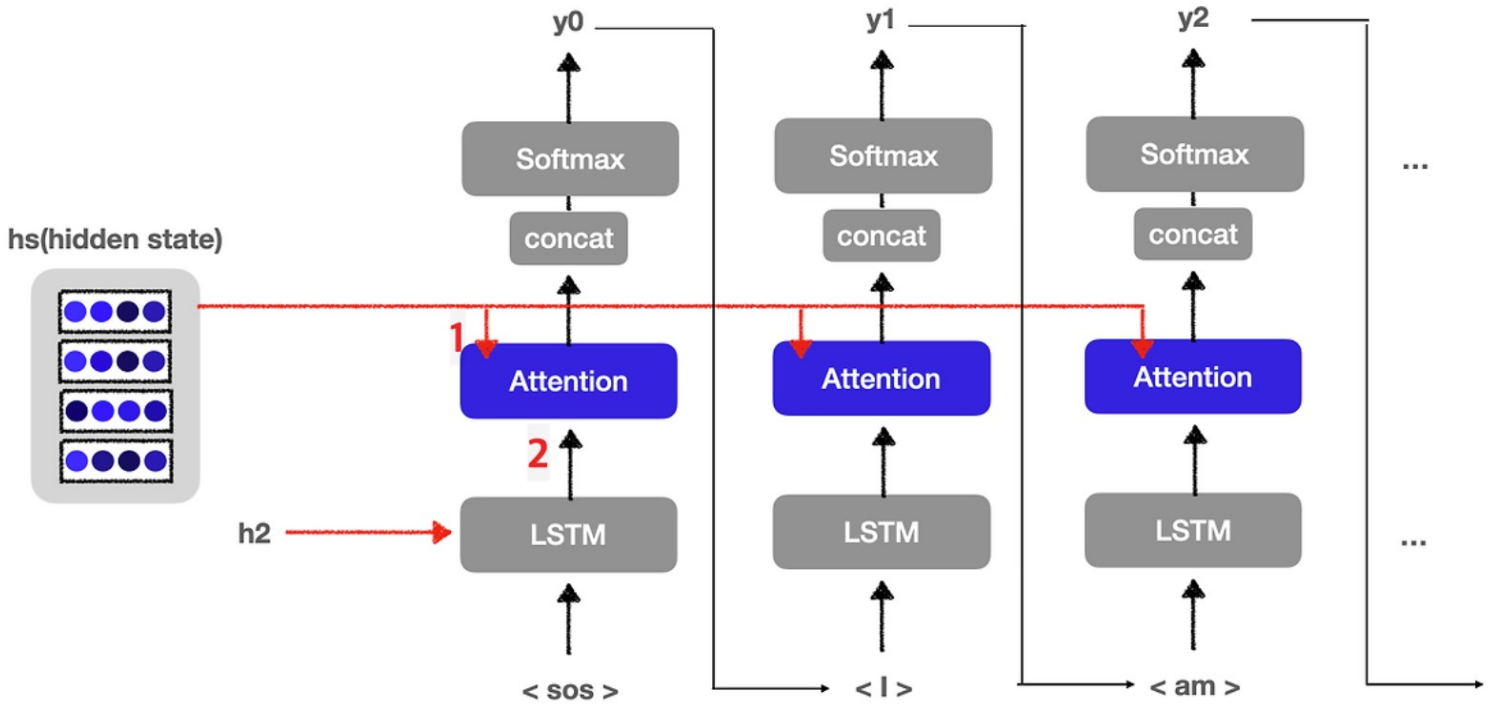


어텐션 메커니즘 작동 예시

2-1. Attention Mechanism에서의 Encoder

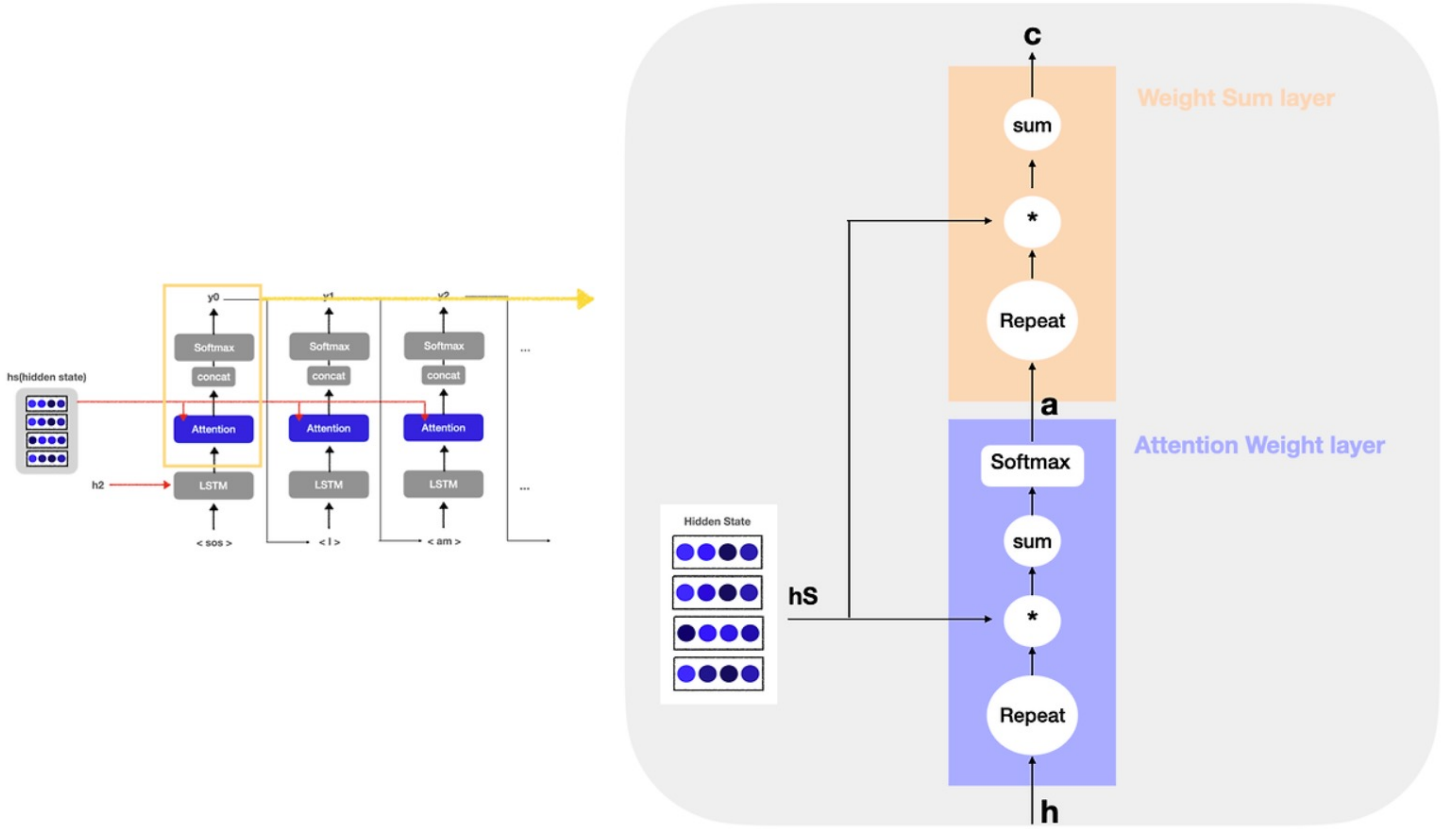


어텐션 메커니즘 작동 예시



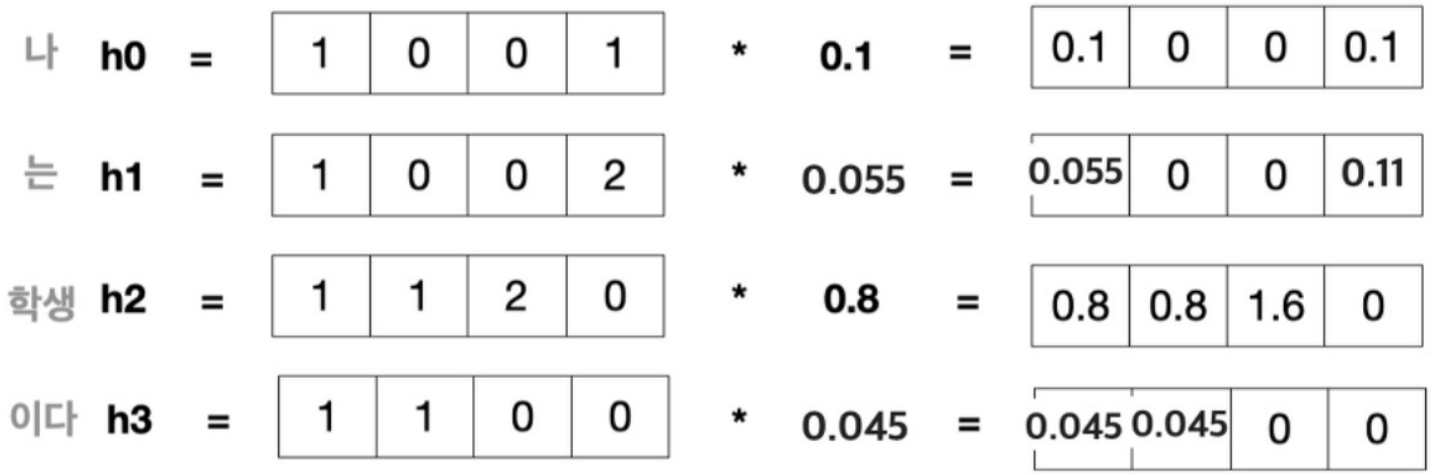
Attention을 포함한 decoder 구조

어텐션 메커니즘 작동 예시



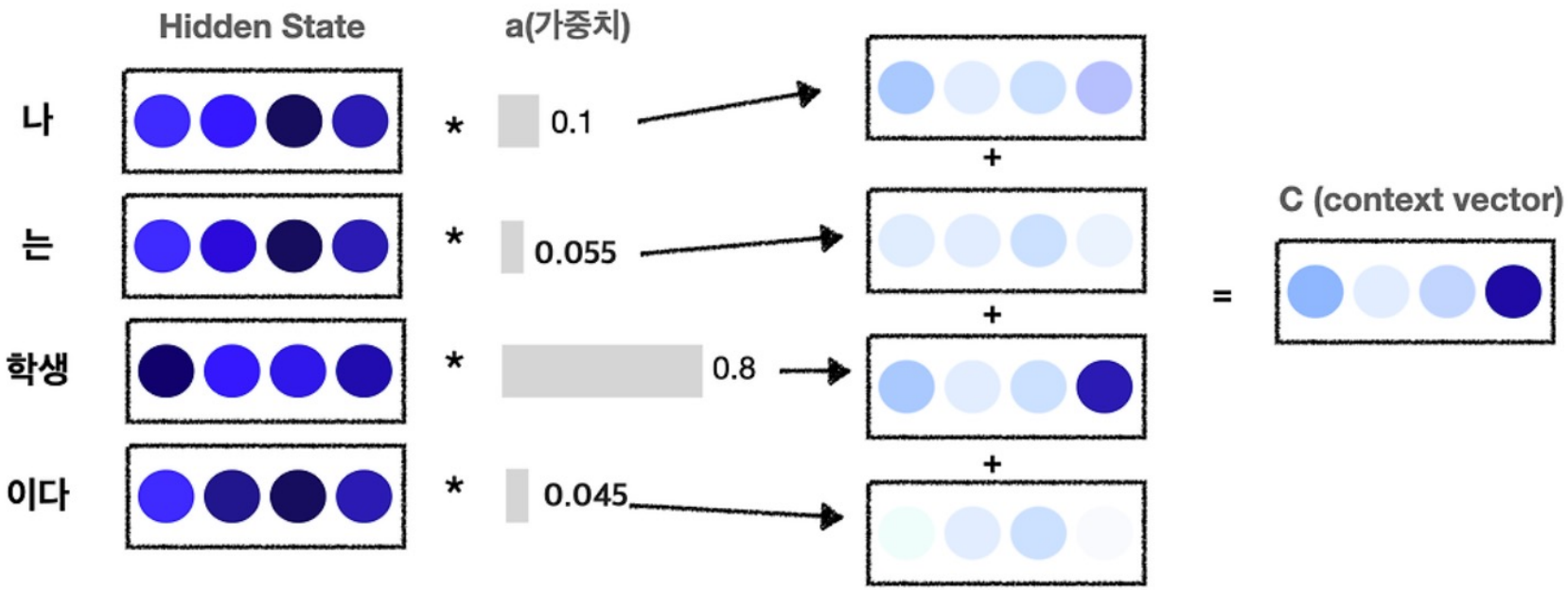
Attention layer

어텐션 메커니즘 작동 예시



이때, Attention Weight layer에서 출력되는 값은 (0.1 0 0 0 0.1) + (0.055 0 0 0.11) + (0.8 0.8 1.6 0) + (0.045 0.045 0 0)

Weight sum layer



각 단어의 가중치와 단어 벡터의 가중합(weighted sum)

Attention 종류

1. 바다나우 어텐션 (Bahdanau Attention)

$$e_{ij} = v_a^\top \tanh(W_a s_{i-1} + U_a h_j)$$

2. 셀프 어텐션 (Self-Attention)
3. Scaled Dot-Production Attention

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

4. 루옹 어텐션 (Luong Attention)
5. 멀티헤드 어텐션 (Multi-Head Attention)

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V$$

Transformer model

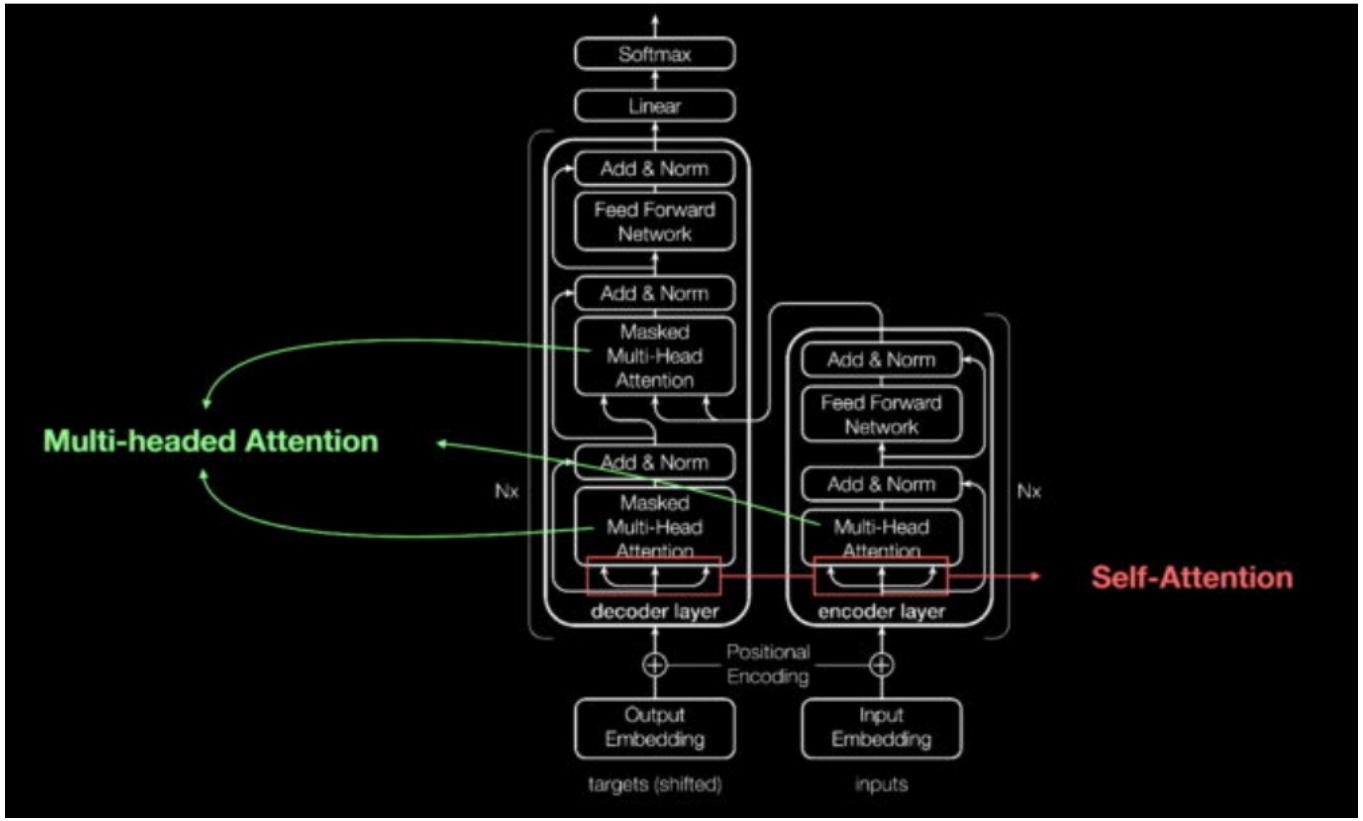
2017 년 Vaswani 등이 발표한 어텐션 메커니즘만을 사용하는 모델

주요특징: 병렬 처리, 확장성, 효율성

Attention is All You Need

- 트랜스포머 모델은 어텐션 메커니즘을 기반으로 한 시퀀스 변환 모델입니다.
- 이 모델은 인코더와 디코더로 구성되어 있으며, 각각 다층의 셀프 어텐션과 feed forward신경망으로 이루어져 있습니다.
- 트랜스포머 순환 신경망 (RNN)이나 합성곱 신경망 (CNN)을 사용하지 않고도 높은 성능을 발휘합니다.
- 이 모델은 병렬화가 용이하고, 학습 속도가 빠르며, 다양한 자연어 처리 작업에서 우수한 성능을 보입니다.

지속적인 연구: 분야의 지속적인 개선과 확신
잠재력: 다양한 AI 및 머신러닝 도메인에서의 넓은 응용 가능성



Transformer에서의 셀프 어텐션 / 멀티헤드 어텐션

설명: 입력 시퀀스의 각 단어가 다른 모든 단어에 주의를 기울여 문맥을 더 효과적으로 이해

멀티헤드 어텐션: 어텐션 메커니즘을 여러 개로 분할하여 서로 다른 부분에 집중.

Attention Based Model

그녀는 주전자의 물을 컵에 따랐다. 그것이 가득 찰 때까지.

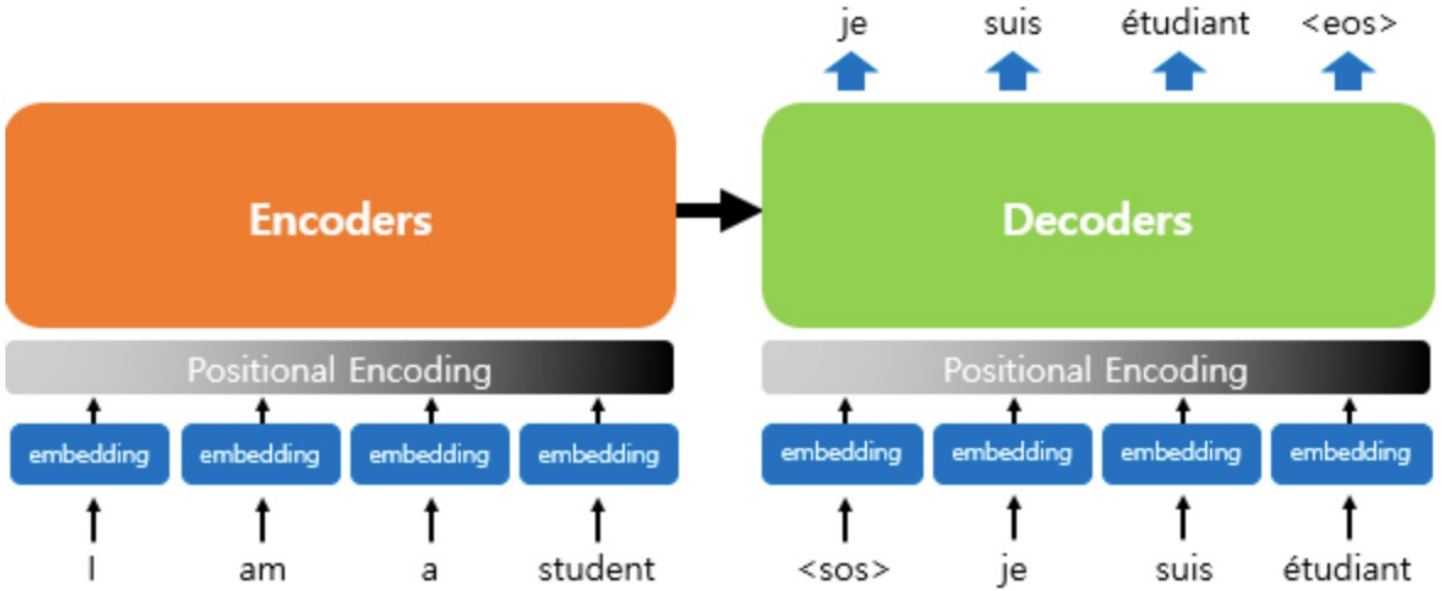
‘그것’ = 컵

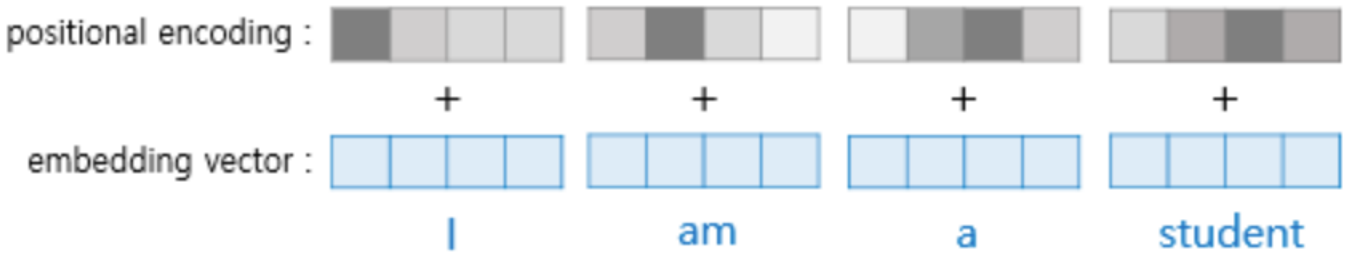
그녀는 주전자의 물을 컵에 따랐다. 그것이 텅 빌 때까지.

‘그것’ = 주전자

Encoder and Decoder stacks

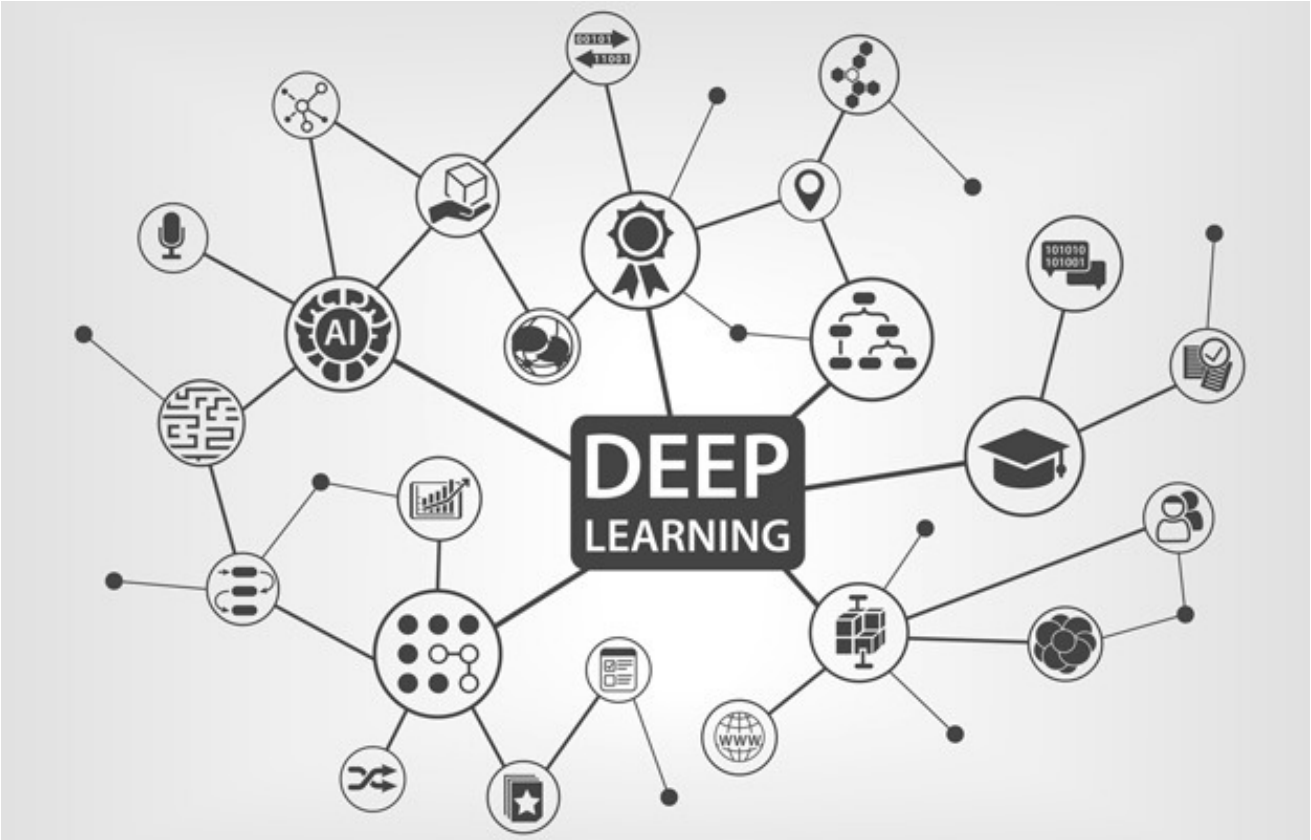
트랜스포머 모델은 위치 인코딩, 멀티-헤드 셀프 어텐션, position wise feed forward 네트워크를 통해 입력 시퀀스의 모든 위치에 주의를 기울여 출력을 생성합니다.



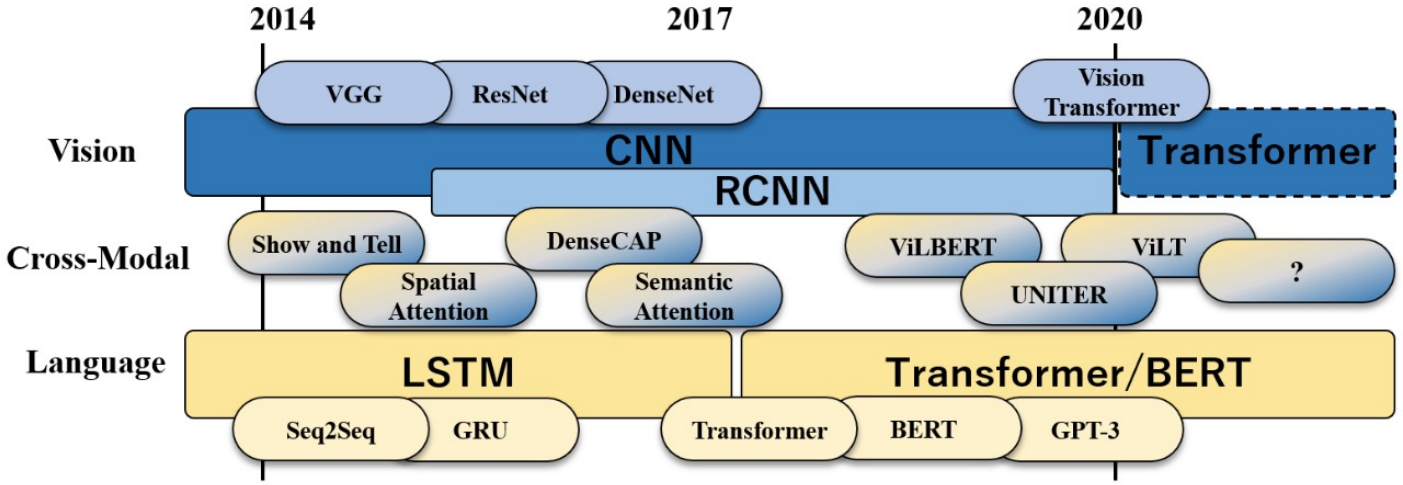


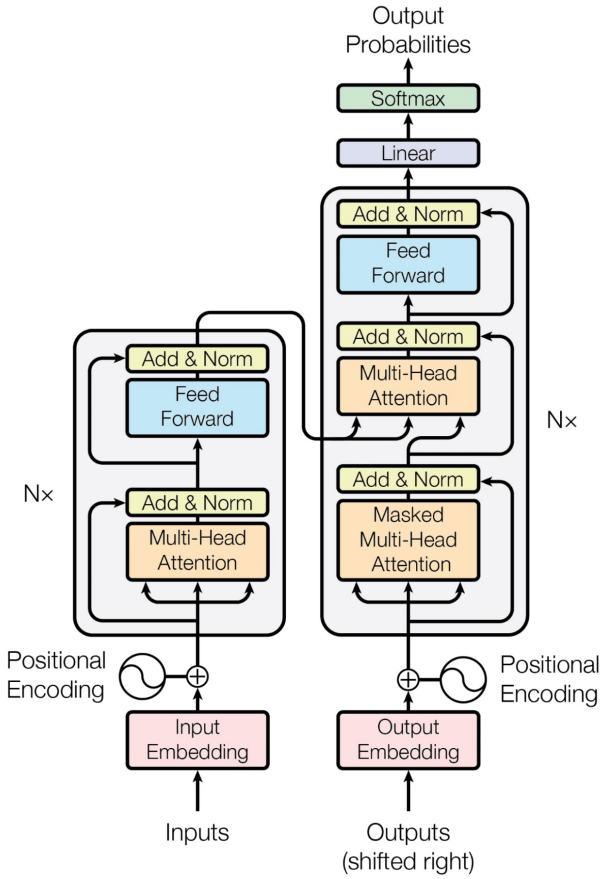
어텐션 메커니즘의 장점

- 성능 향상: 언어 이해와 생성에서 정확도 향상
- 문맥 인식: 시퀀스 내의 장거리 의존성을 더 잘 처리
- 확장성: 대규모 데이터셋의 효율적인 처리



Transformer vs Traditional Models





결론

중요성: 딥러닝과 어텐션 메커니즘이 NLP와 그 외 분야에 미치는 변혁적인 영향 강조

Model	BLEU		Training Cost (in FLOPS * 10^{18})	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet (Kalchbrenner et al., 2016)	23.75			
Deep-Att + PosUnk (Zhou et al., 2016)		39.2		100
GNMT + RL (Wu et al., 2016)	24.6	39.92	23	140
ConvS2S (Gehring et al., 2017)	25.16	40.46	9.6	150
MoE (Shazeer et al., 2017)	26.03	40.56	20	120
GNMT + RL Ensemble (Wu et al., 2016)	26.30	41.16	180	1100
ConvS2S Ensemble (Gehring et al., 2017)	26.36	**41.29**	77	1200
Transformer (base model)	27.3	38.1	**3.3**	
Transformer (big)	**28.4**	**41.8**	23	

Future Diretions