Natural Language Processing
- RNN 부터 BERT까지 -

참고: 토크ON세미나 https://www.youtube.com/watch?v=qlxrXX5uBoU&t=343s

목차

자연어 처리

자연어와 자연어 처리의 개념을 살펴보고, 자연어 처리의 방법 중 하나인 Word Embedding에 대해서 살펴보겠습니다.

1. 자연어란?

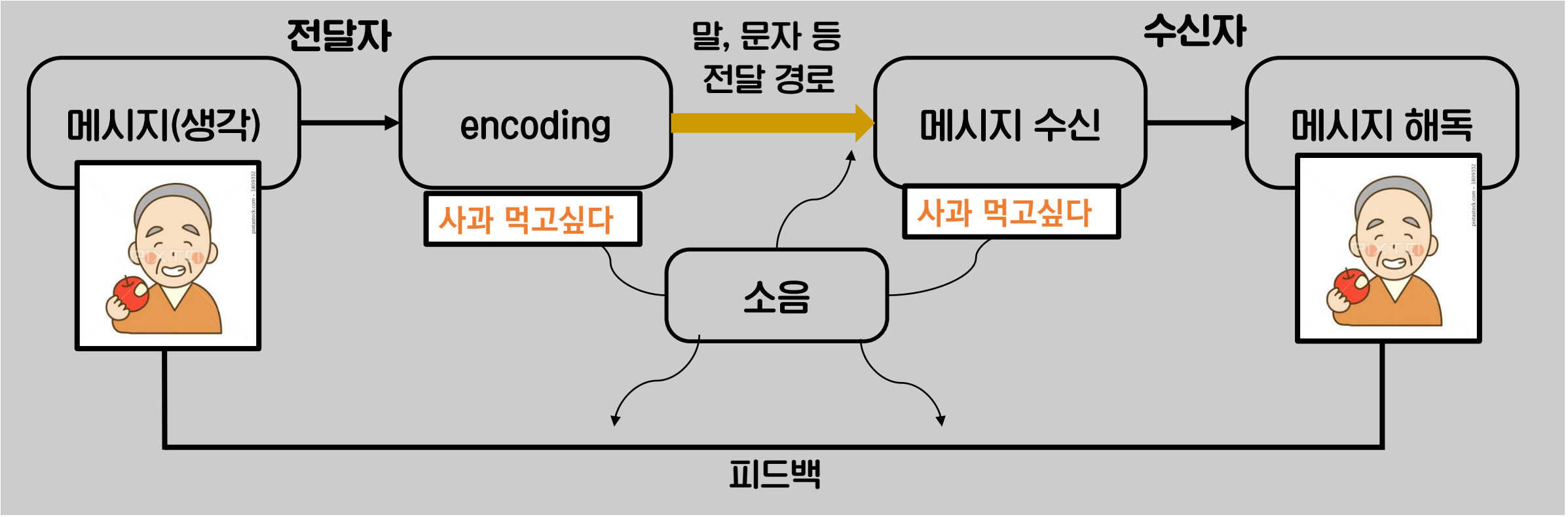
자연어의 정의

자연 언어: 일반 사회에서 자연히 발생하여 쓰이는 언어 (각국의 언어: 한국어, 영어, 일본어, ...)

인공 언어: 의도적으로 만들어진 언어 (JAVA, Python, ...)

자연어의 정의

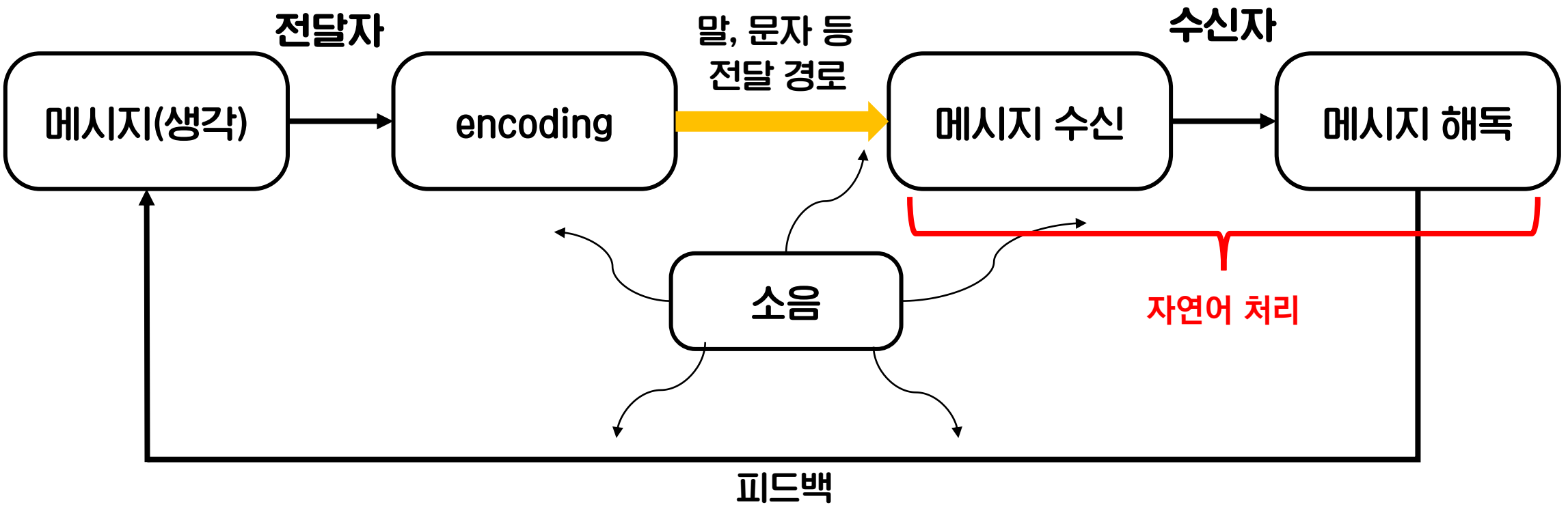
자연어 처리: 컴퓨터를 이용하여 인간 언어의 이해, 생성 및 분석을 다루는 인공지능 기술



1. 자연어란?

자연어의 정의

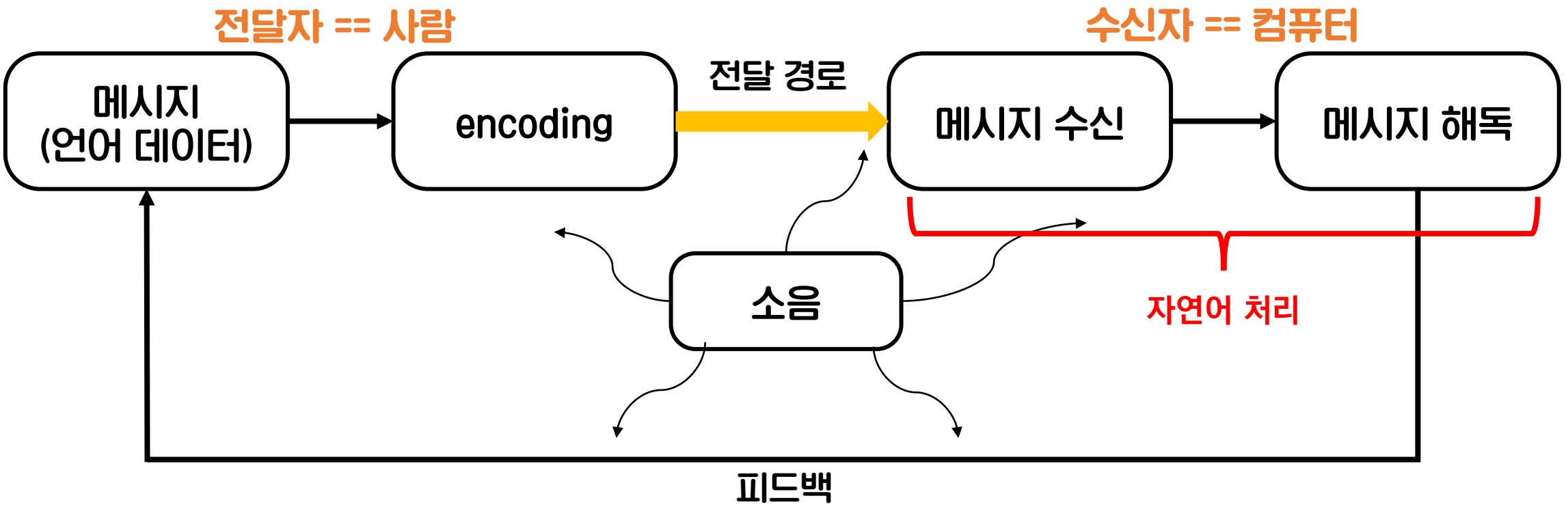
자연어 처리: 컴퓨터를 이용하여 인간 언어의 이해, 생성 및 분석을 다루는 인공지능 기술



1. 자연어란?

자연어의 정의

자연어 처리: 컴퓨터를 이용하여 인간 언어의 이해, 생성 및 분석을 다루는 인공지능 기술



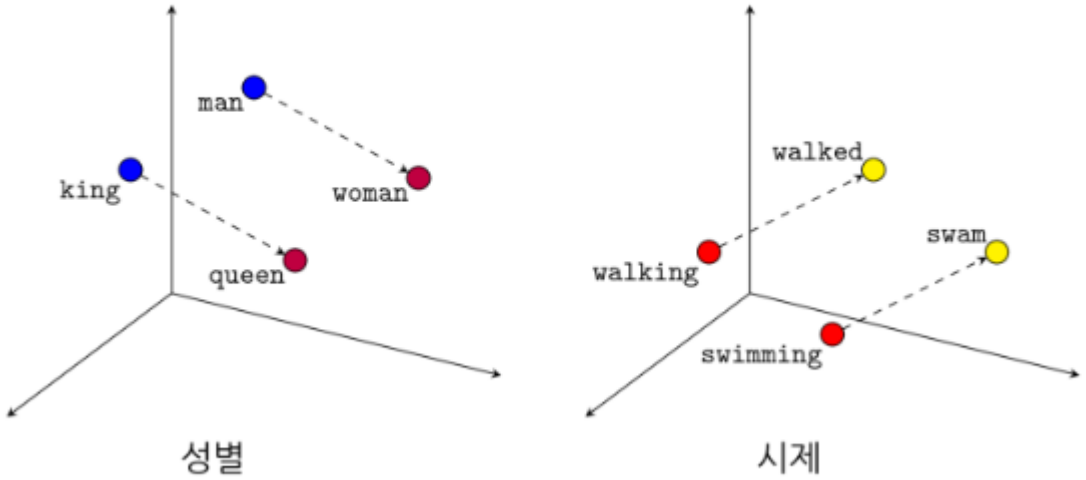
Word Embedding이란?

word embedding: 자연어를 어떻게 수학적으로 설명하느냐.

- 컴퓨터가 볼 때, 언어는 '기호'로 보인다. => 컴퓨터는 자연어의 의미를 모름.

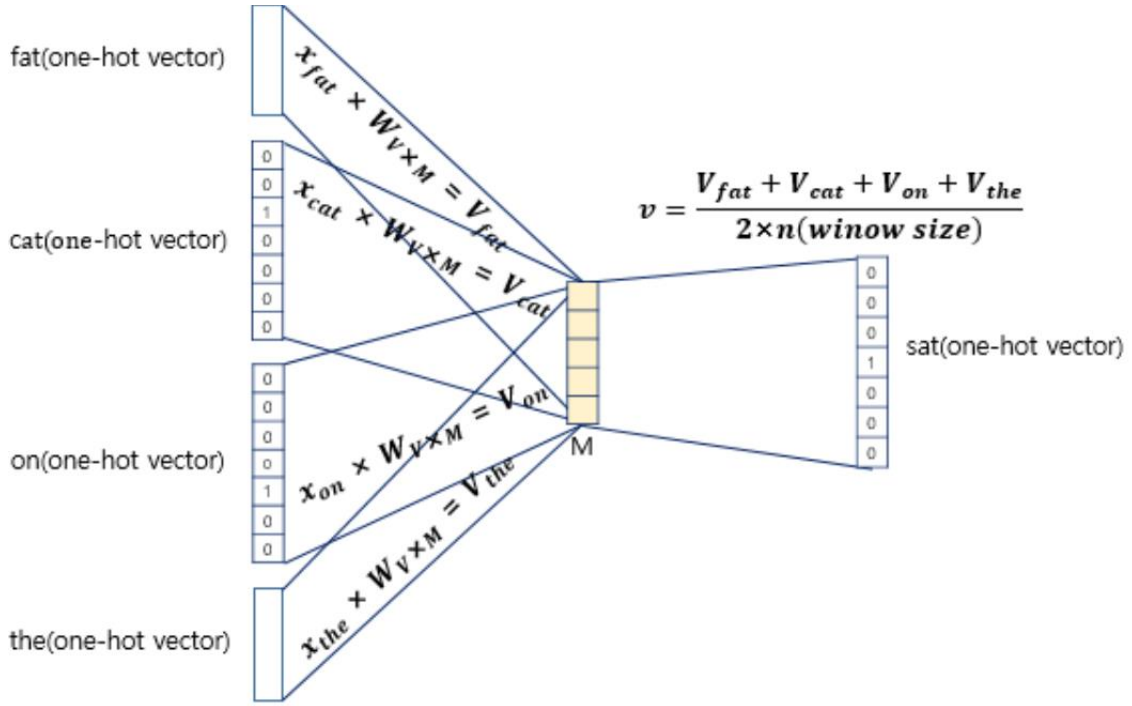
Word2Vec: 자연어의 '의미'를 벡터공간에 임베딩

- How? : 주변 단어를 통해 의미를 임베딩



Word Embedding이란?

word	one-hot-vector
The	[1, 0, 0, 0, 0, 0, 0]
fat	[0, 1, 0, 0, 0, 0, 0]
cat	[0, 0, 1, 0, 0, 0, 0]
sat	[0, 0, 0, 1, 0, 0, 0]
on	[0, 0, 0, 0, 1, 0, 0]
the	[0, 0, 0, 0, 0, 1, 0]
mat	[0, 0, 0, 0, 0, 0, 1]



언어 모델

언어 모델의 정의를 알아보고,
다양한 언어 모델들을 살펴보겠습니다.

언어 모델 이란?

모델 이란?

- 대상 주제를 수식이나 기호를 사용하여 표현한 것.
- 자연의 법칙을 컴퓨터로 시뮬레이션하거나 미래의 상태(state)를 예측하도록 학습시킴.

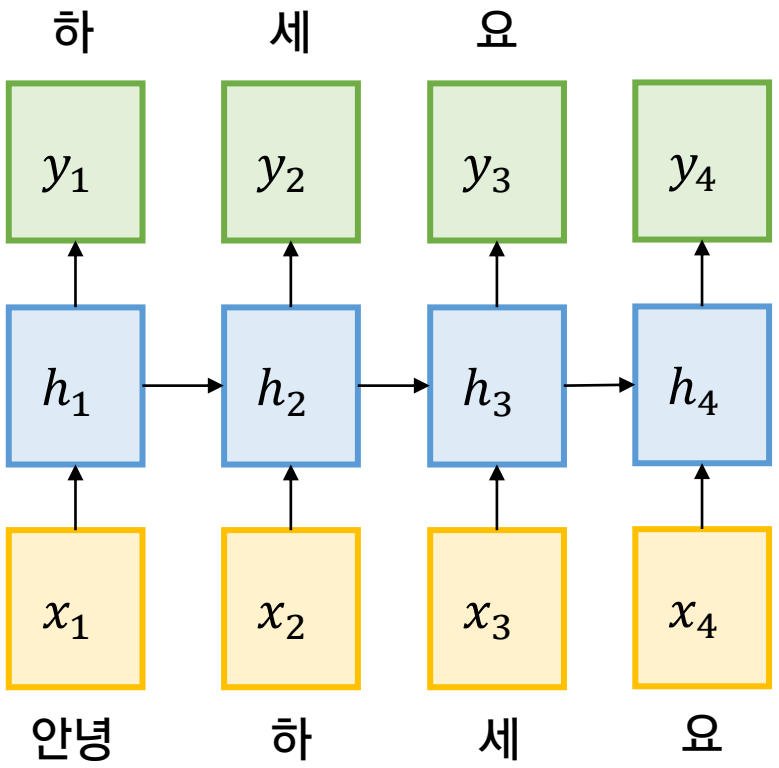
ex) 일기예보모델, 데이터 모델, ...

- 원하는 작업을 수행해주는 프로그램

언어 모델이란?

- '자연어'의 법칙을 컴퓨터로 모사한 모델
- 주어진 단어들로 다음 나올 단어 예측 (이전 state 로부터 다음 state 를 예측)
- 다음 등장 단어를 잘 예측하는 모델은 그 언어의 특성을 잘 반영한 모델이자 문맥을 잘 계산한 모델.

RNN 모델



x_t : t시점의 input

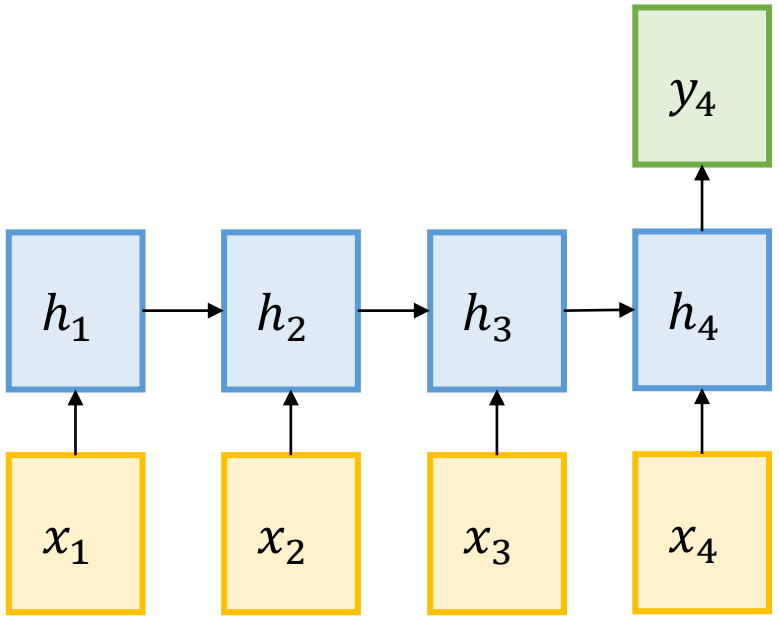
h_t : t시점의 hidden state

x_t 가 x_{t+1} 을 예측하도록 학습

y_t : t시점의 output

x_t와 h_{t-1}, 그리고 h_t 를 사용하여 x_{t+1} 예측

RNN 모델



마지막 출력은 앞선 단어들의 문맥을 고려해서 생성된 출력.

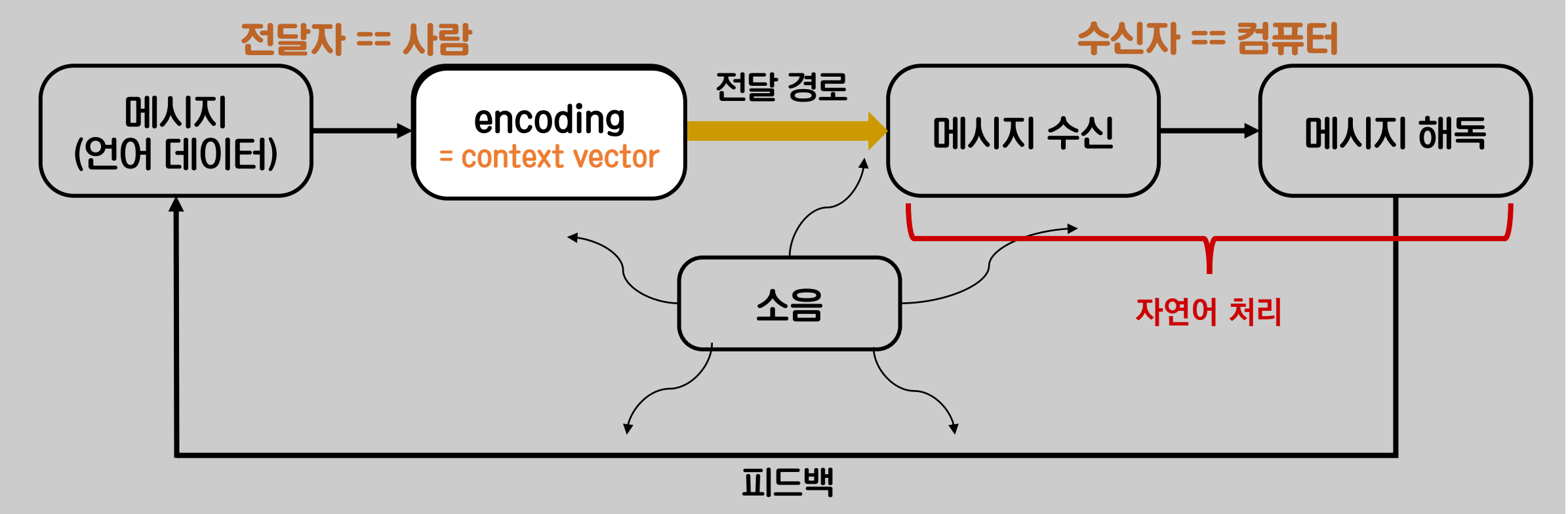
문맥정보를 학습하여 얻은 context vector

=> 자연어의 문맥을 인코딩 가능!

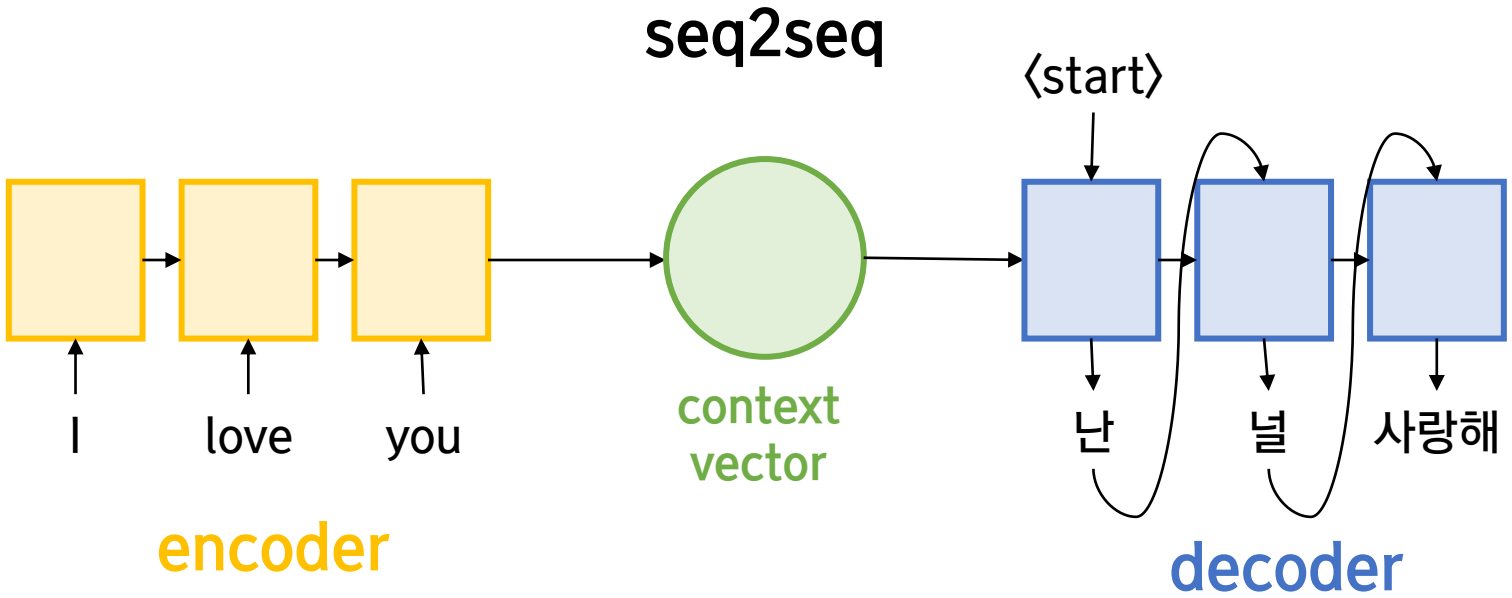
=> 문맥정보를 활용하여 자연어 처리 가능!

자연어의 정의

자연어 처리: 컴퓨터를 이용하여 인간 언어의 이해, 생성 및 분석을 다루는 인공지능 기술



RNN 모델



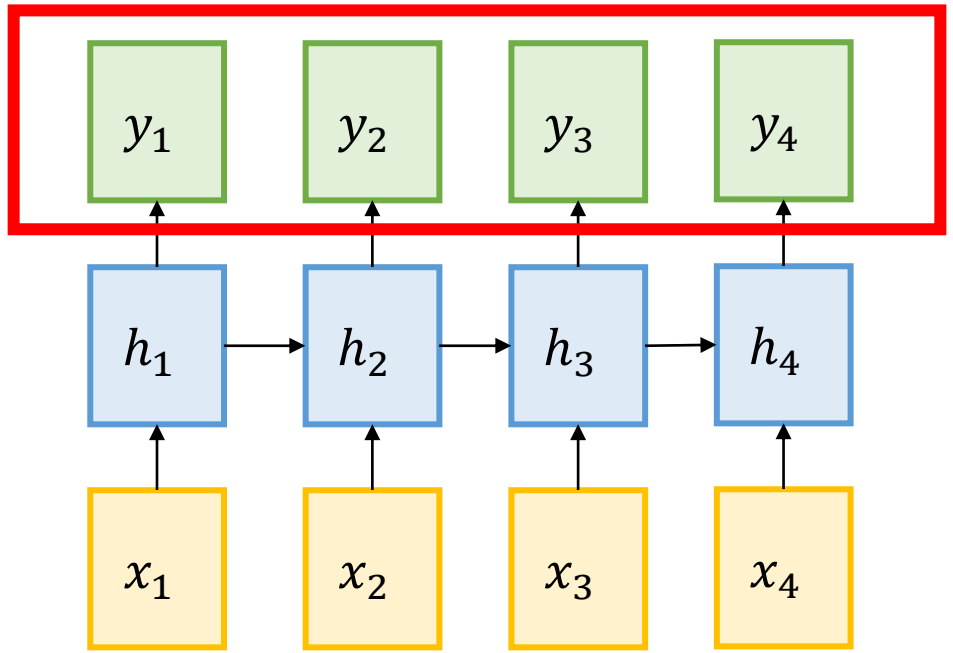
단점

- 문장의 길이가 길면 토큰의 의미가 희석됨.
- 고정된 context vector 크기로 인해 긴 문장의 의미를 나타내기 어려움.
- 모든 토큰이 영향을 미쳐서 필요 없는 단어들도 영향을 똑같이 줌.

→ Attention 모델 제안

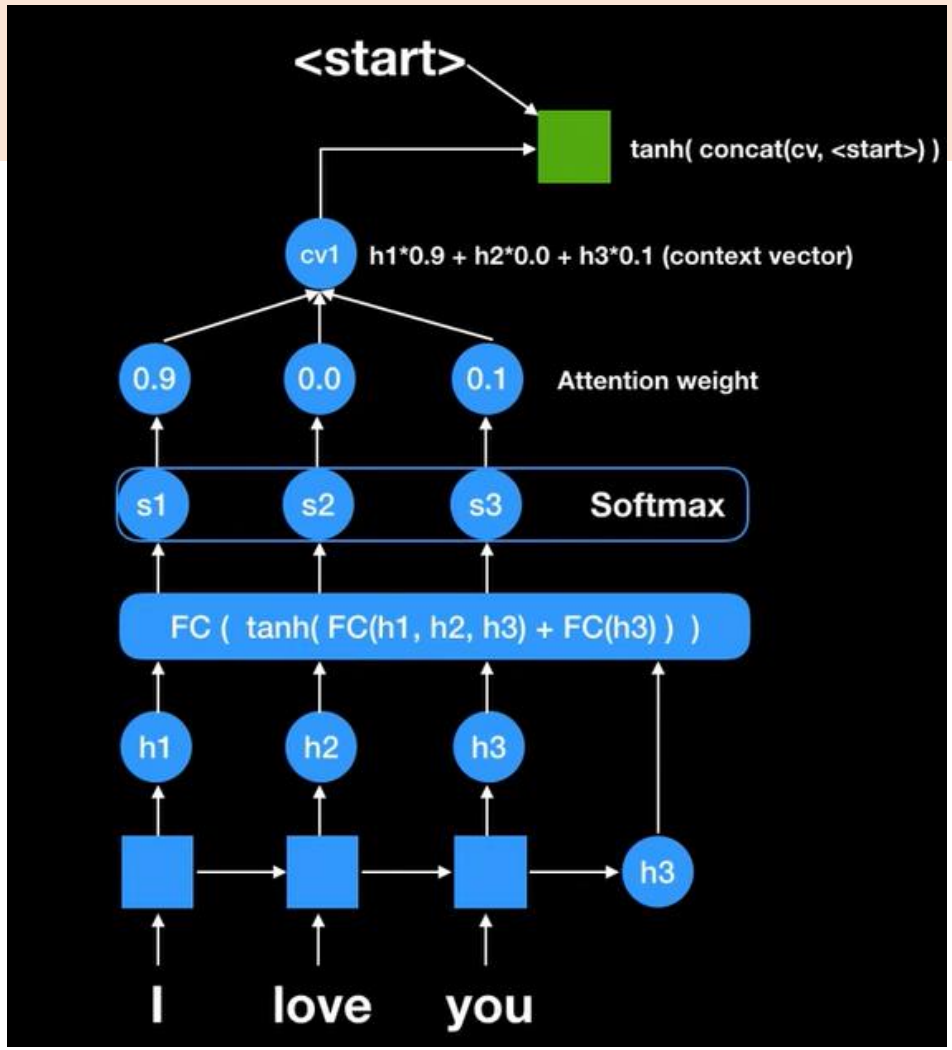
Attention 모델

- 인간이 정보처리할 때 모든 정보에 집중하지 않고, 중요한 정보에 집중함.
- 중요한 토큰을 중요하게 다루자!

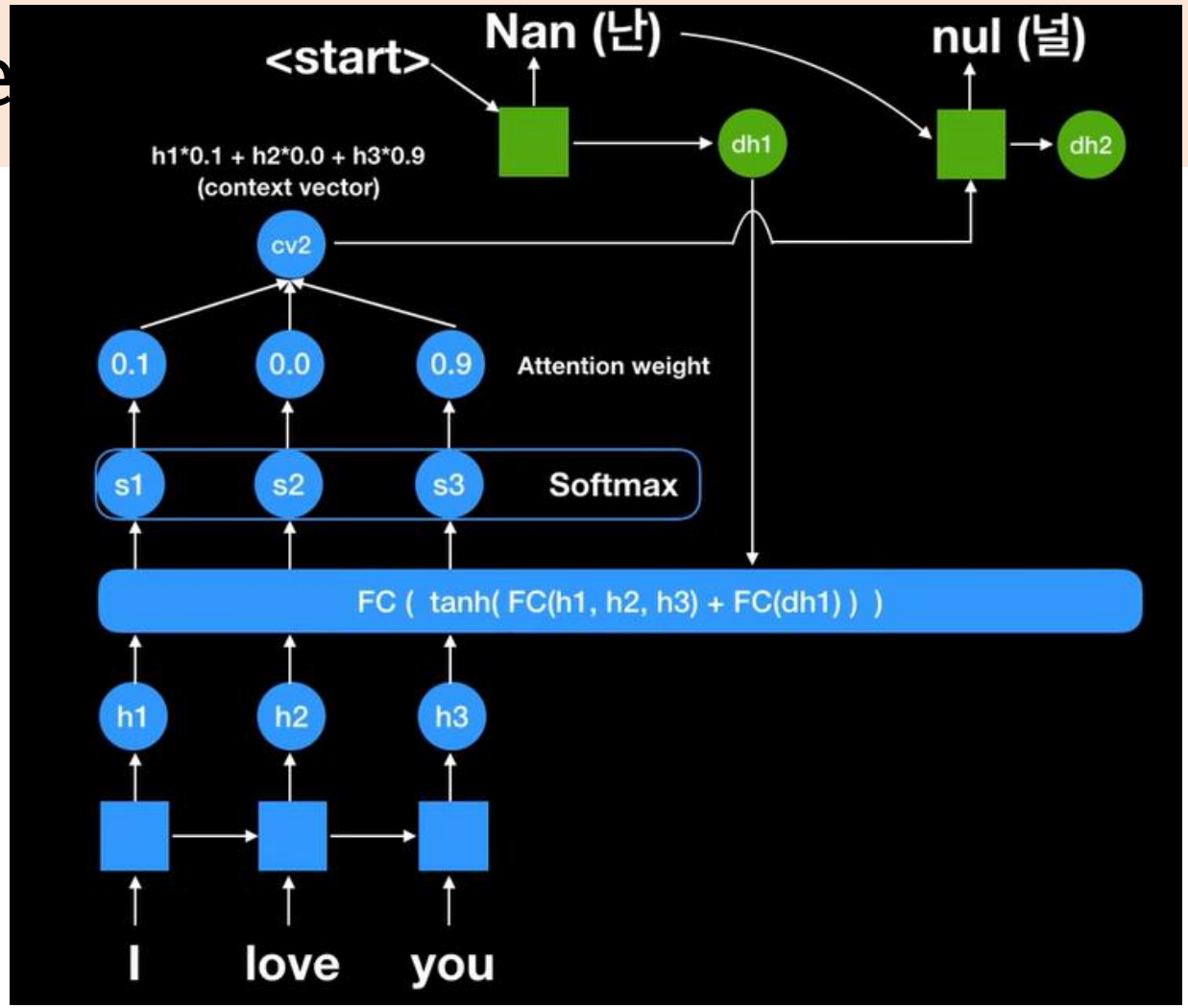


=〉 이 정보를 유용하게 활용해보자!

언어 모델 | 3. Attention 모델

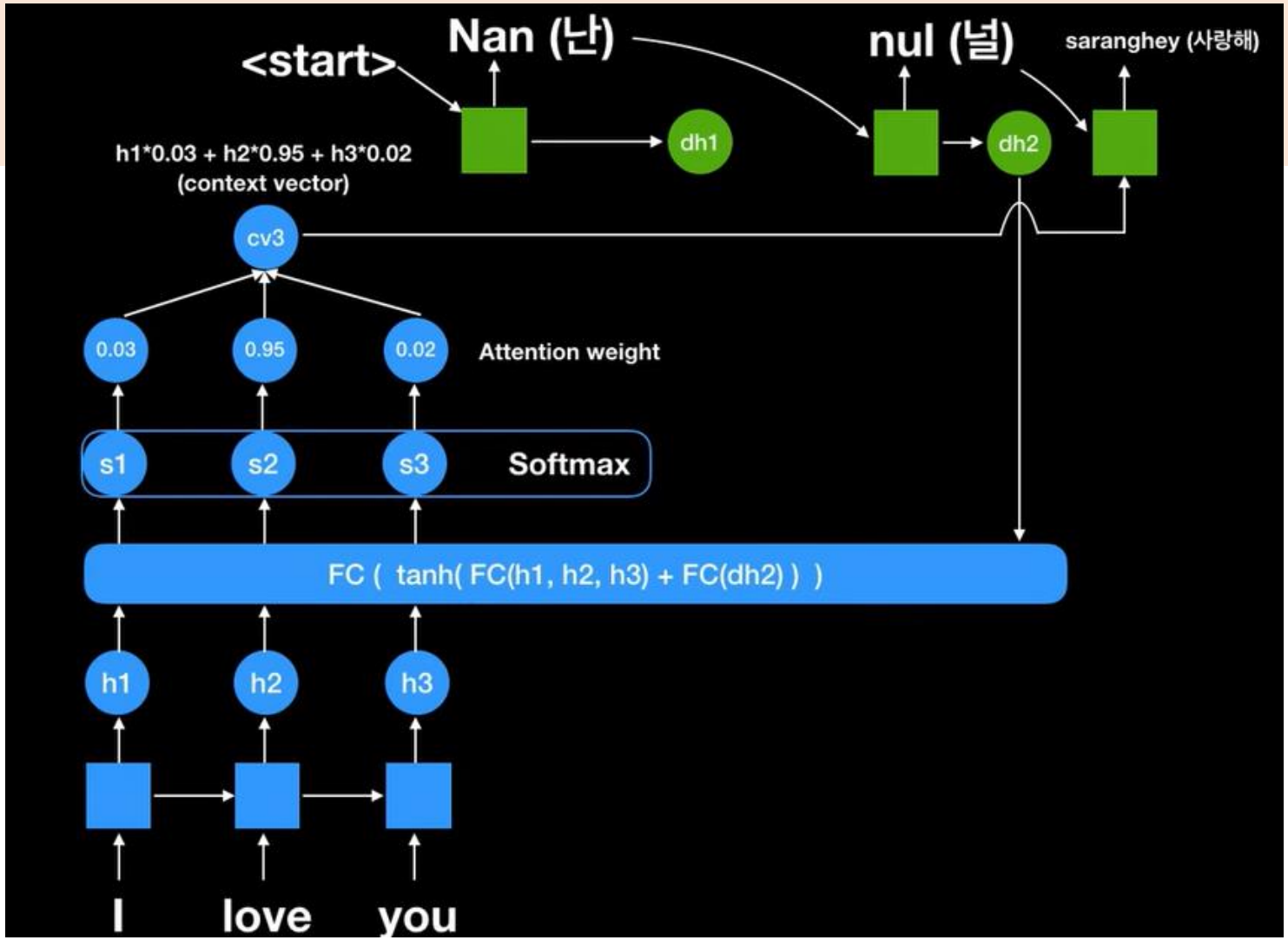


Atte



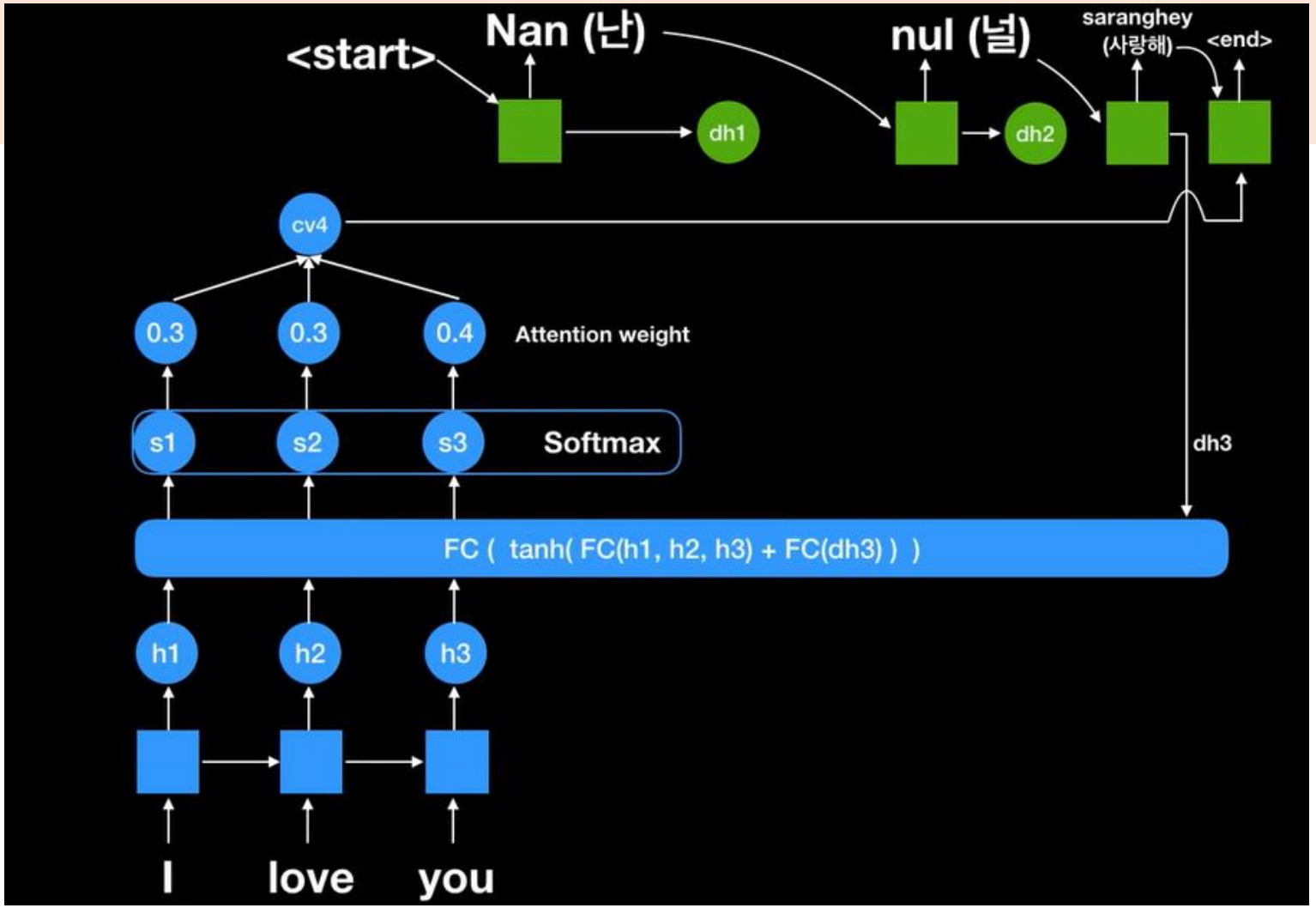
https://www.youtube.com/watch?v=WsQLdu2JMgI

언어 모델 | 3. Attention 모델



https://www.youtube.com/watch?v=WsQLdu2JMgI

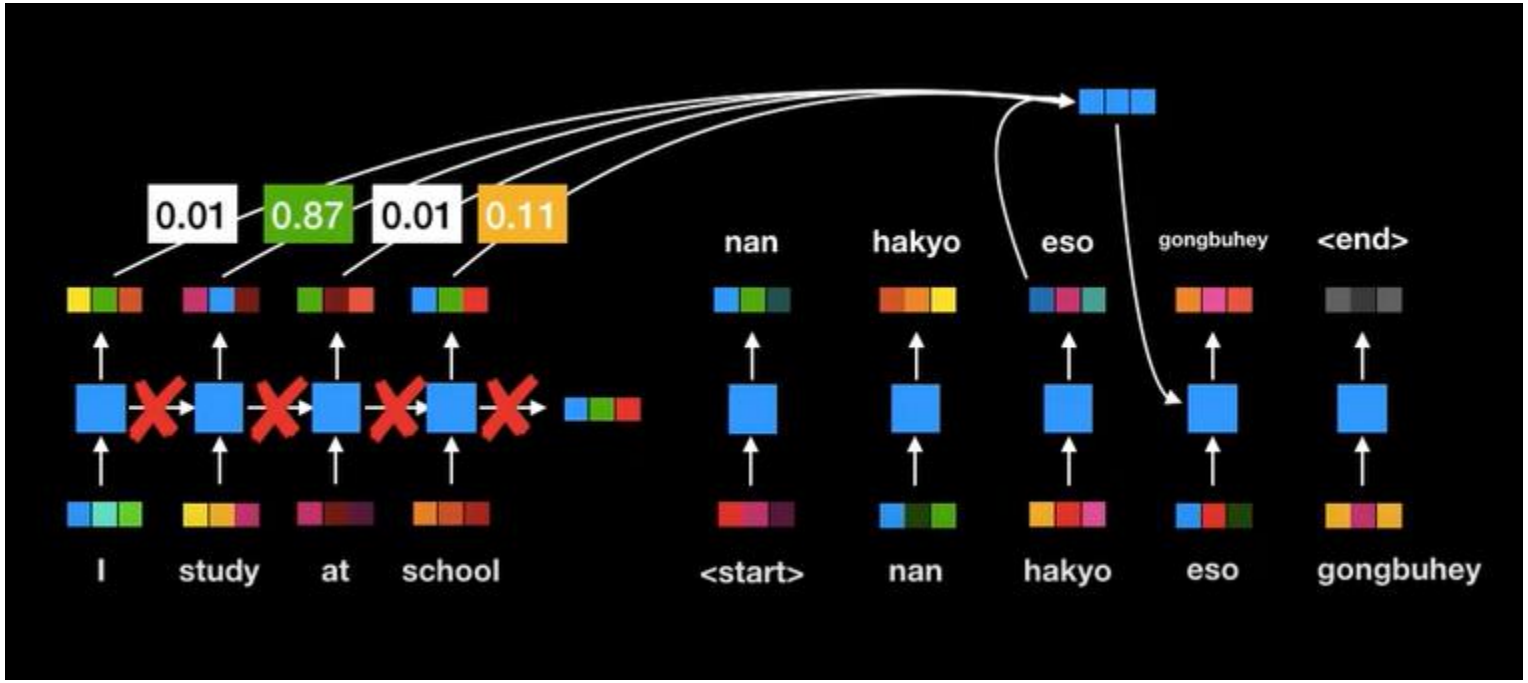
언어 모델 | 3. Attention 모델



https://www.youtube.com/watch?v=WsQLdu2JMgI

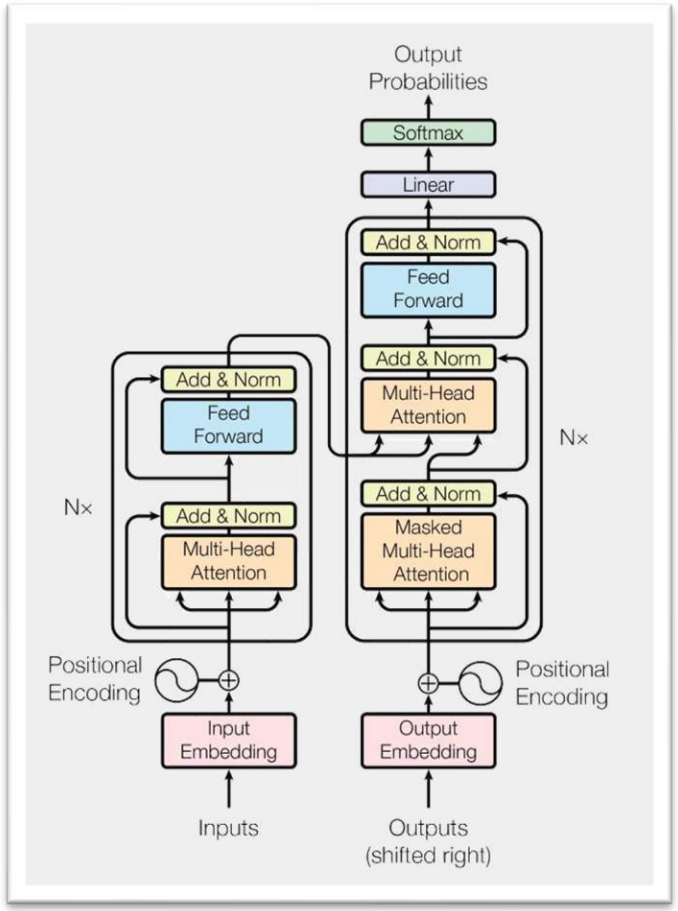
3. Attention 모델

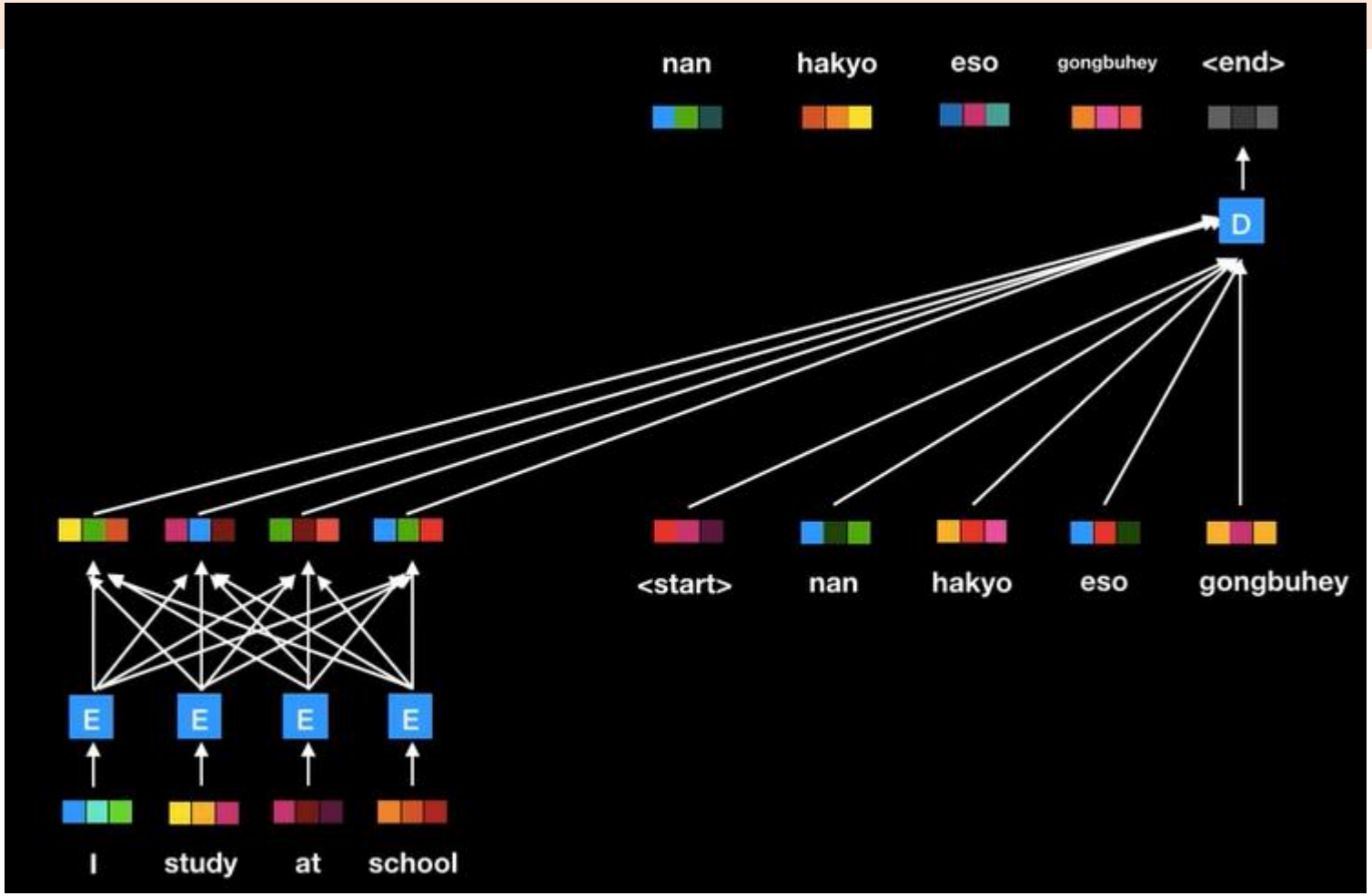
Self-Attention



3. Attention 모델

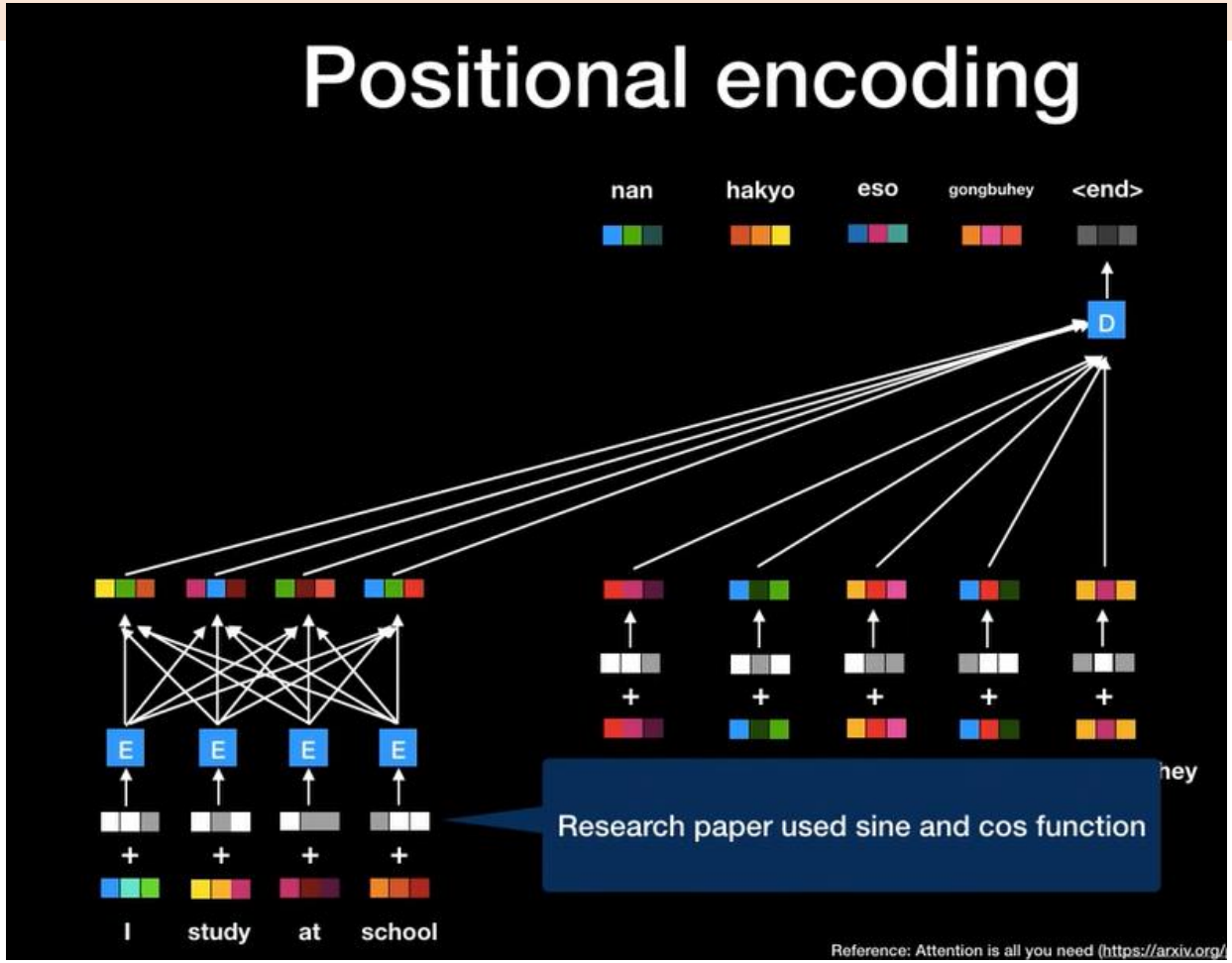
Self-Attention



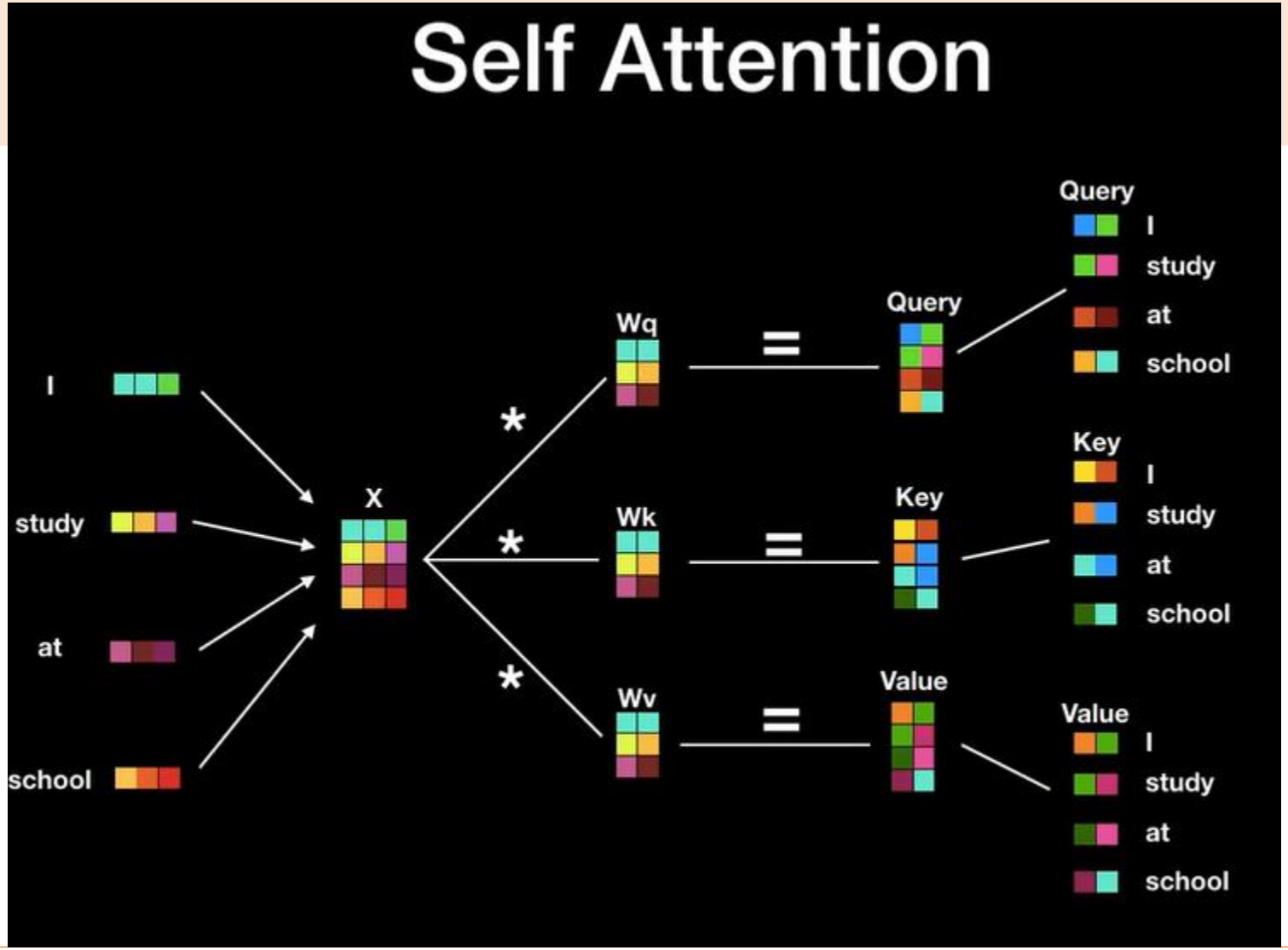


3. Attention 모델

Self-Attention



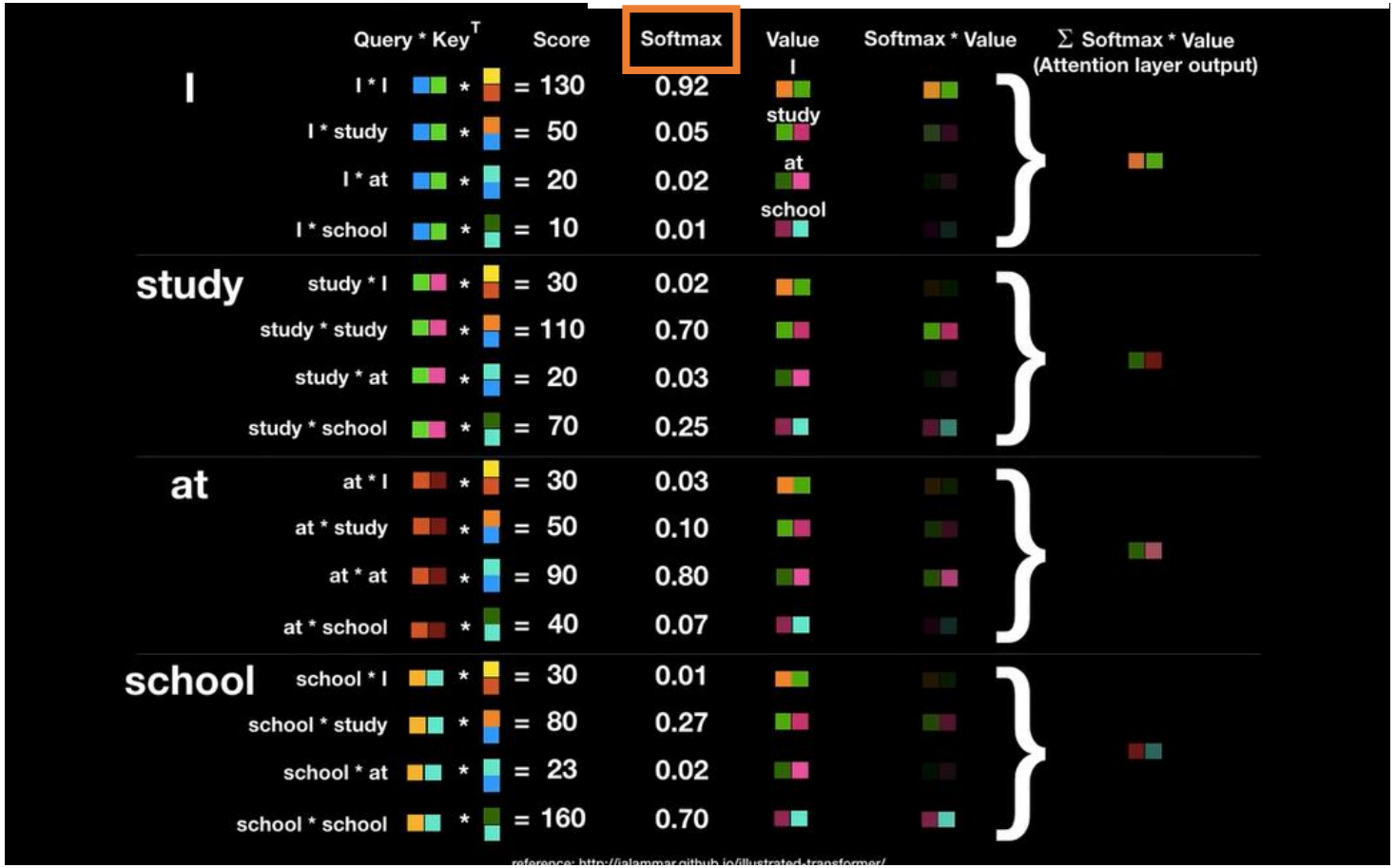
3. Attention 모델



3. Attention 모델

Self-Attention

key가 query와 어느정도 연관성이 있는지



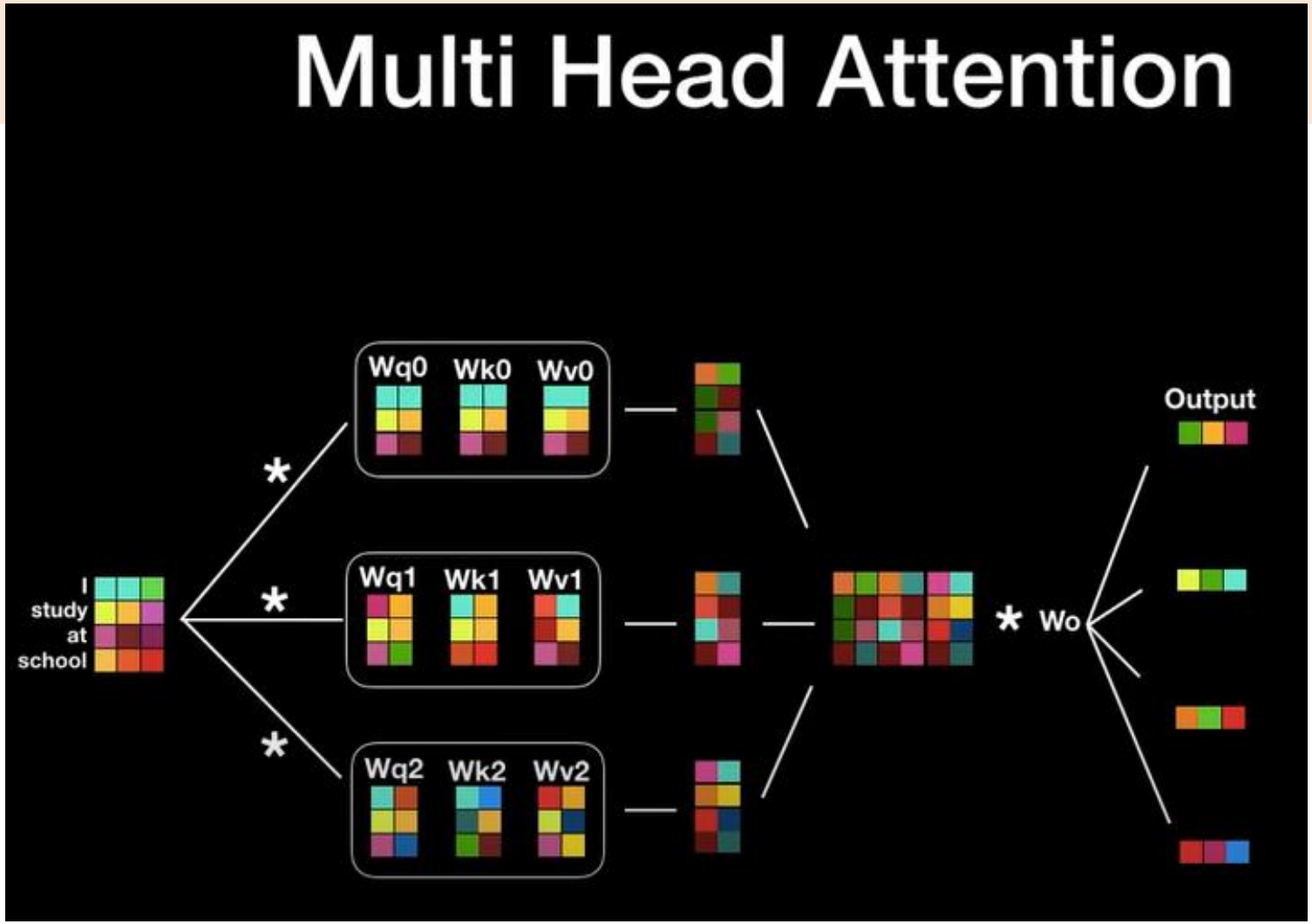
Query: 현재 단어

Key: Query와의 상관관계를 나타내고자 하는 대상

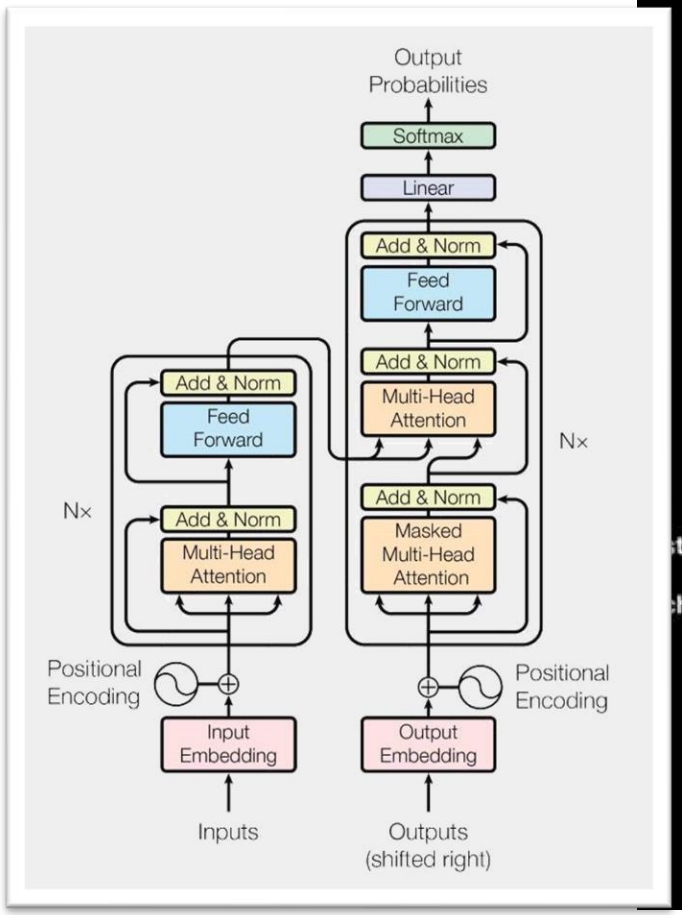
Attention score = Query * key

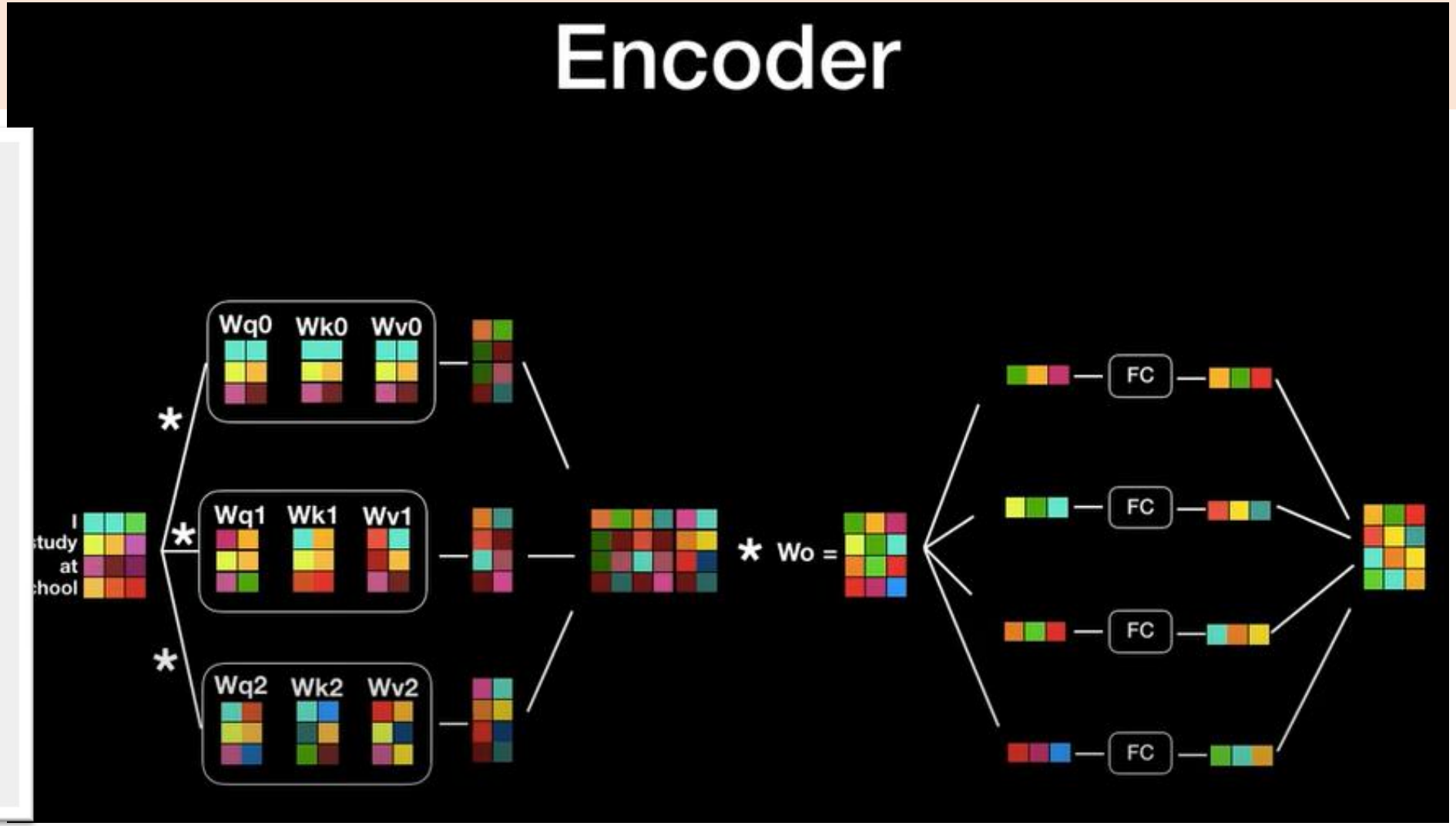
value: Query와 key의 상관관계 값

3. Attention 모델

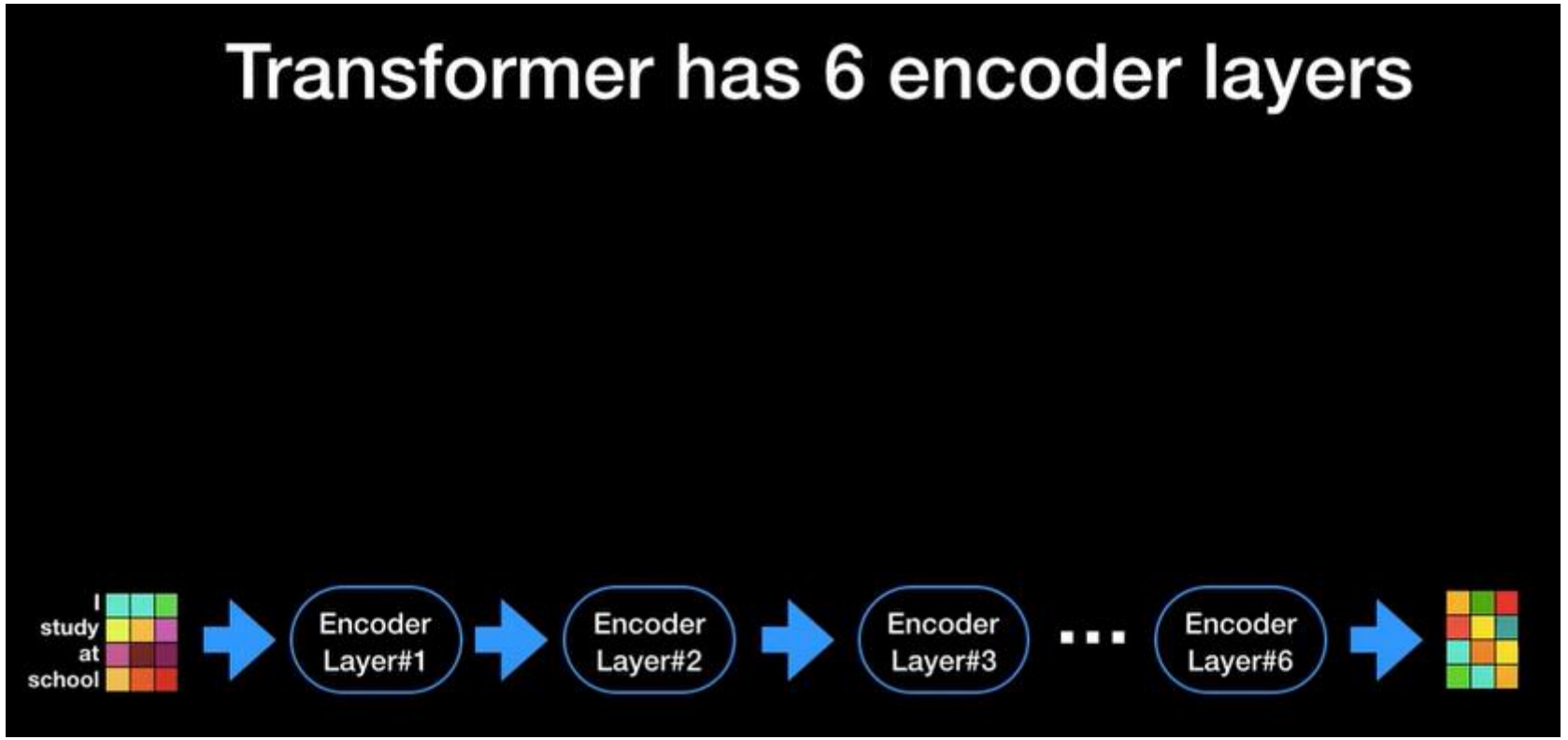


언어 모델 | 3. Attention 모델



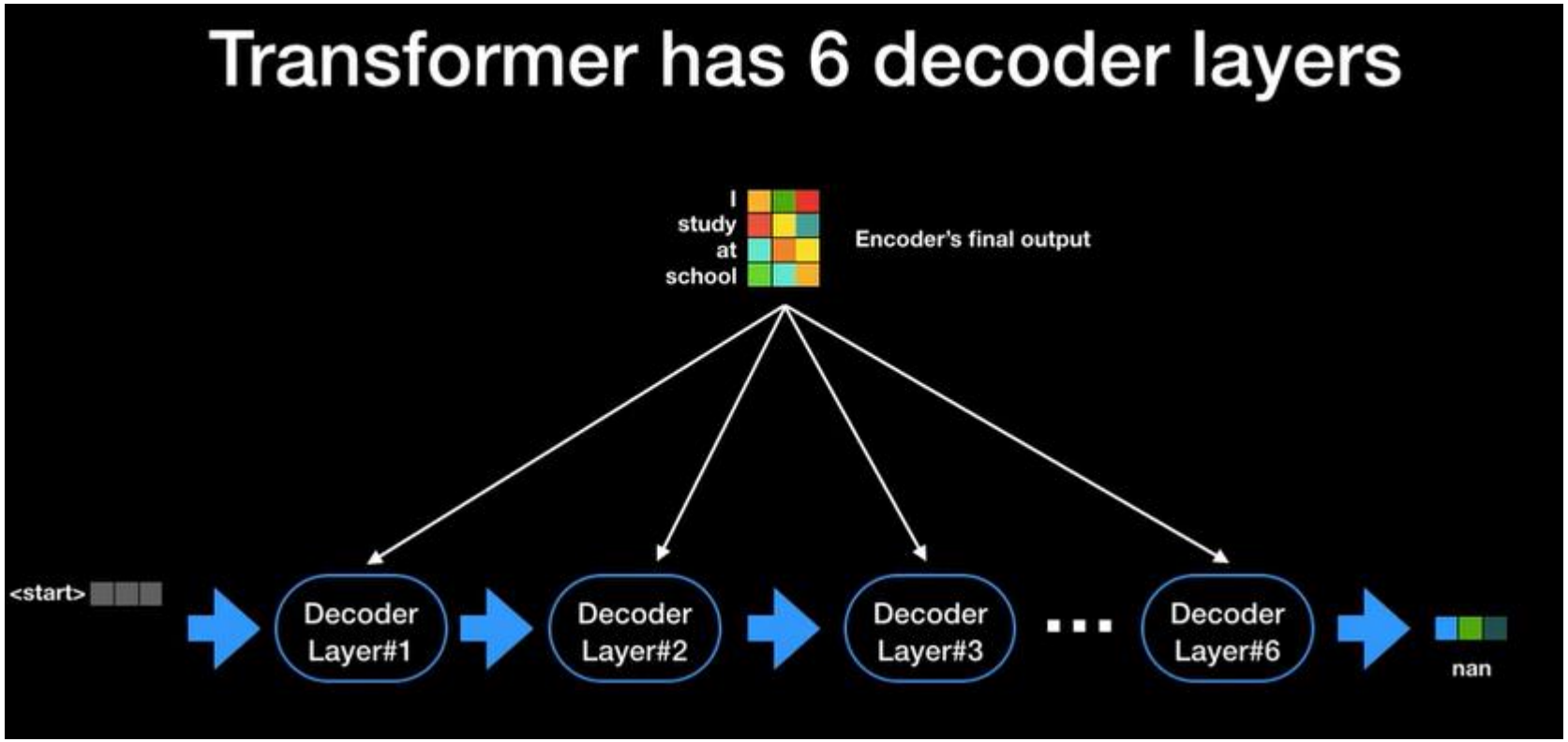


Transformer 모델

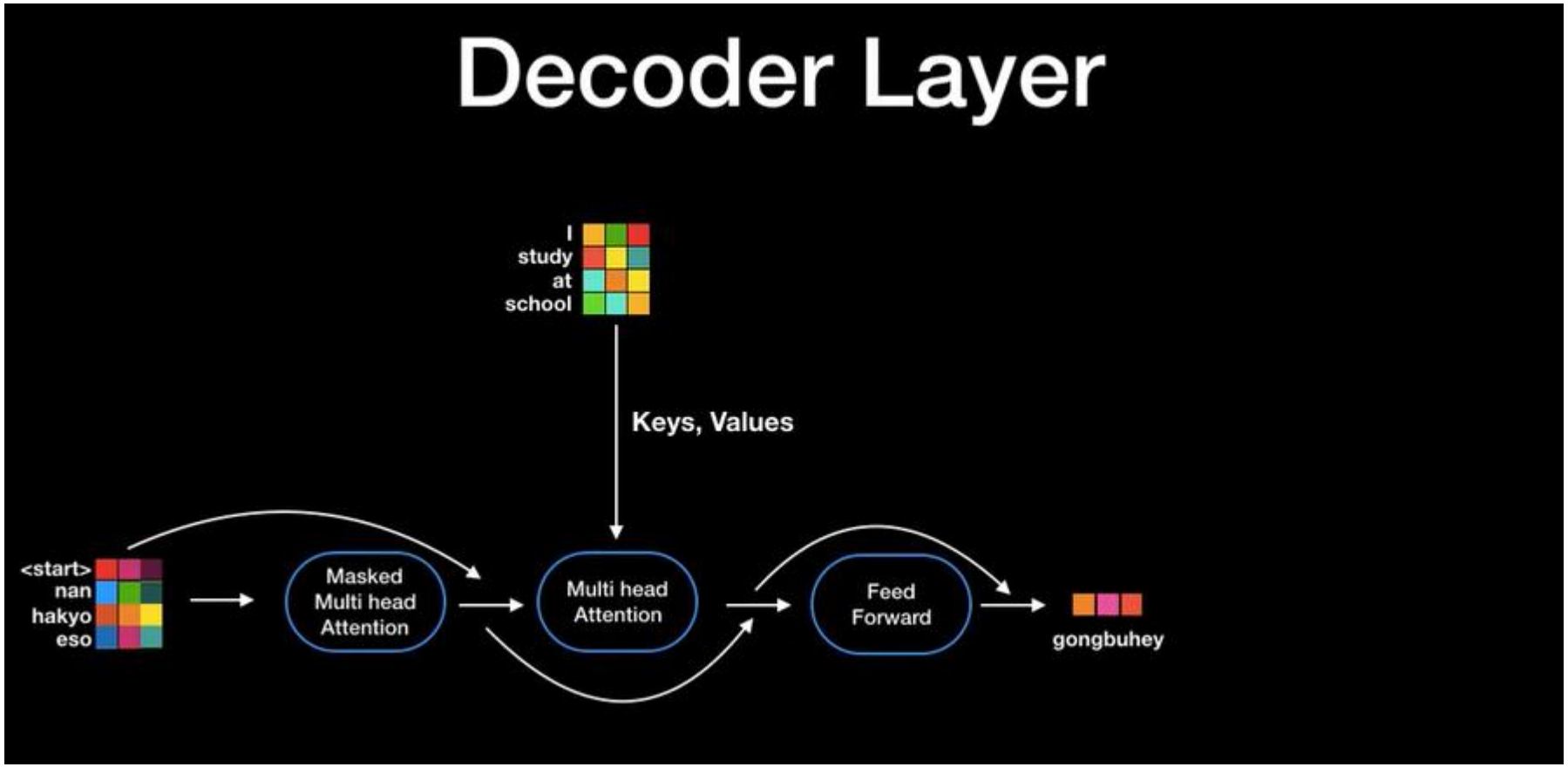


4. Transformer 모델

Transformer 모델



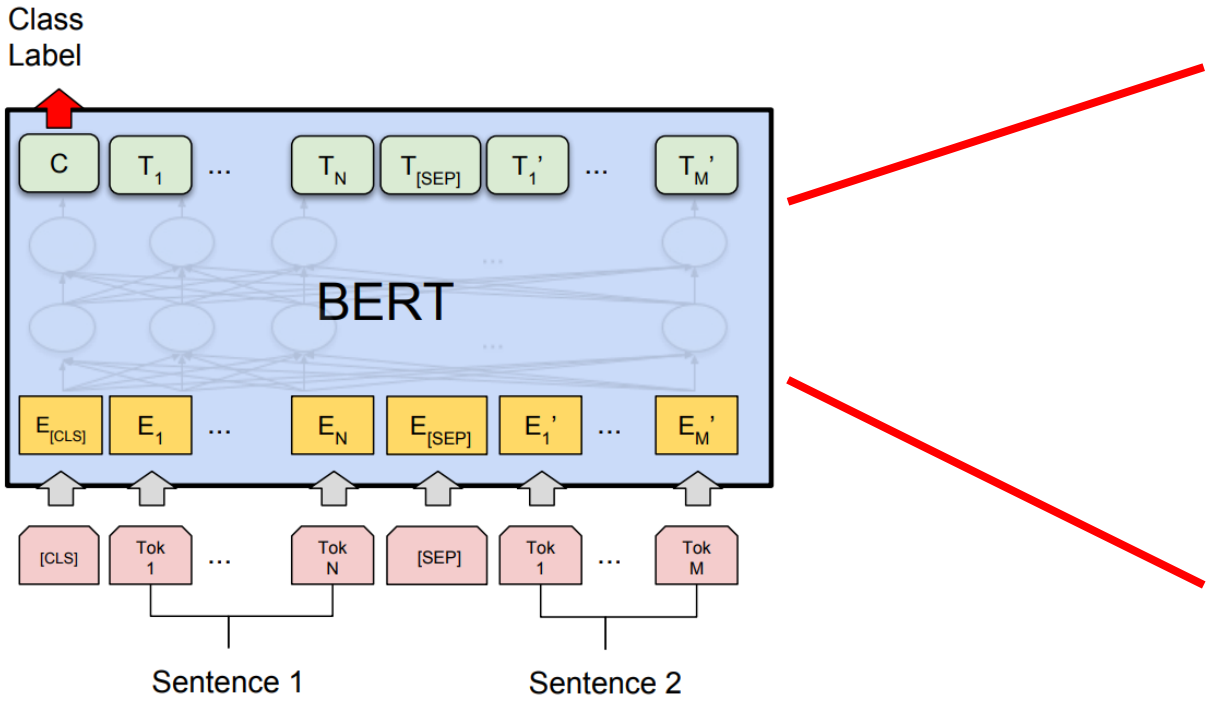
Transformer 모델



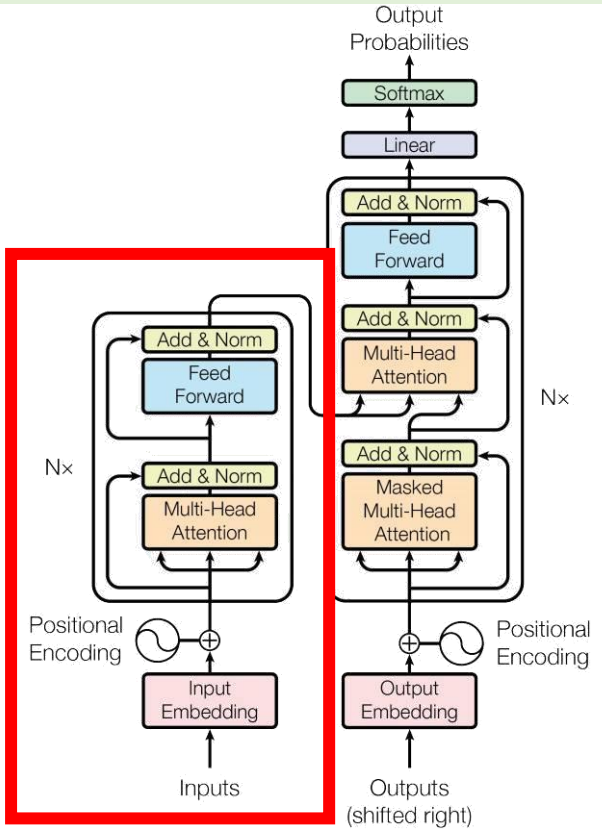
BERT

자연어 처리에서 가장 널리 쓰이는
BERT모델에 대해서 알아보겠습니다.

BERT 모델 구조



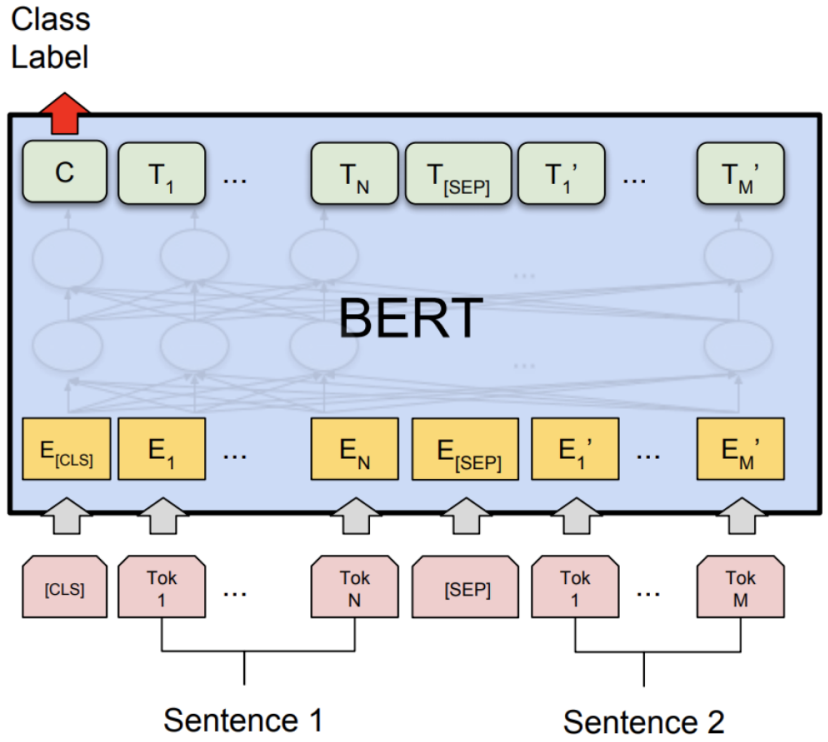
BERT



Transformer

BERT의 구조

BERT 모델



Contextual Representation Token

Transformer Layer (Encoder layer)

Input Embedding Layer

1. WordPiece Tokenizer

WordPiece Tokenizer

- Byte Pair Encoding(BPE) 알고리즘 사용
- 빈도수에 기반하여 단어를 의미 있는 패턴(Subword)로 잘라서 tokenizing
- 자주 등장하는 단어 -> 그 자체로 token
- 자주 등장하지 않는 단어 -> 더 잘게 쪼갠 subword token

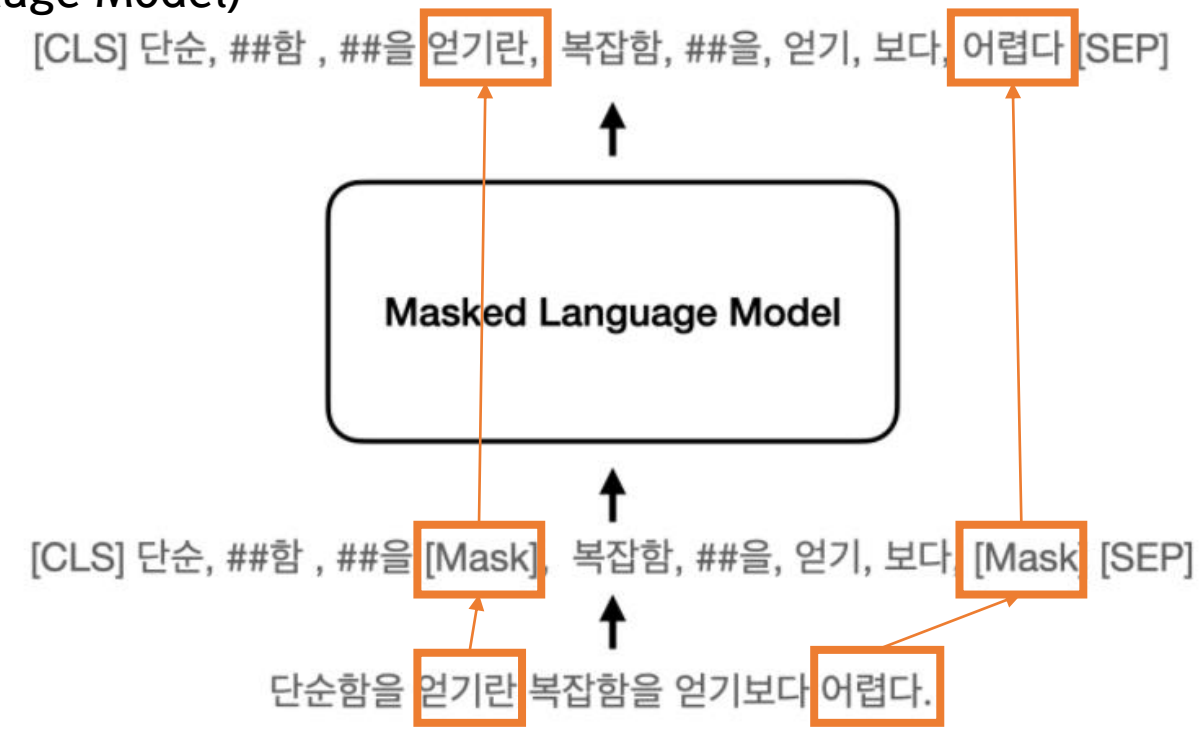
```
tokenize("This is the Hugging Face course!")

>>> ['Th', '##i', '##s', 'is', 'th', '##e', 'Hugg', '##i', '##n', '##g', 'Fac', '##e',
'c', '##o', '##u', '##r', '##s',
 '##e', '[UNK]']
```

https://huggingface.co/course/chapter6/6?fw=pt

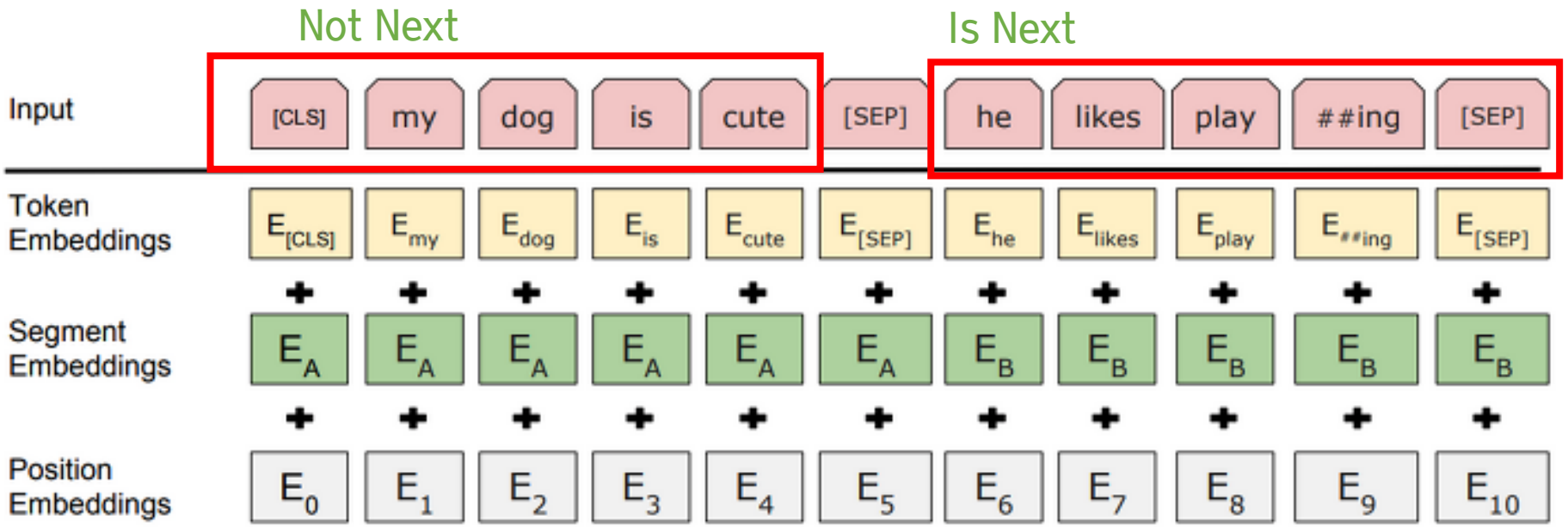
BERT 모델 - 사전학습(MLM)

MLM(Masked Language Model)



BERT 모델 – 사전학습(NSP)

NSP(Next Sentence Prediction)



BERT 응용 모델

1. BERT (from Google) released with the paper BERT: Pre-training of Deep Bi Transformers for Language Understanding by Jacob Devlin, Ming-Wei Char Kristina Toutanova.
2. GPT (from OpenAI) released with the paper Improving Language Understa Pre-Training by Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Su
3. GPT-2 (from OpenAI) released with the paper Language Models are Unsup Learners by Alec Radford, *Jeffrey Wu*, Rewon Child, David Luan, Dario Amo Sutskever**.
4. Transformer-XL (from Google/CMU) released with the paper Transformer- Language Models Beyond a Fixed-Length Context by Zihang Dai, *Zhilin Yang* Carbonell, Quoc V. Le, Ruslan Salakhutdinov.
5. XLNet (from Google/CMU) released with the paper XLNet: Generalized Autoregressive Pretraining for Language Understanding by Zhilin Yang, *Zihang Dai*, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le.
6. XLM (from Facebook) released together with the paper Cross-lingual Language Model Pretraining by Guillaume Lample and Alexis Conneau.
7. RoBERTa (from Facebook), released together with the paper a Robustly Optimized BERT Pretraining Approach by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov.
8. DistilBERT (from HuggingFace), released together with the blogpost Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT by Victor Sanh, Lysandre Debut and Thomas Wolf.

참고 자료

- SK 토크ON세미나 자연어 언어모델 BERT(1~4): https://www.youtube.com/watch?v=qlxrXX5uBoU&t=343s
- HuggingFace Wordpiece Tokenization course: https://huggingface.co/course/chapter6/6?fw=pt
- Youtube seq2seq~attention 설명 : https://www.youtube.com/watch?v=WsQLdu2JMgI&t=528s
- Youtube Transformer 설명: https://www.youtube.com/watch?v=mxGCEWOxfe8