

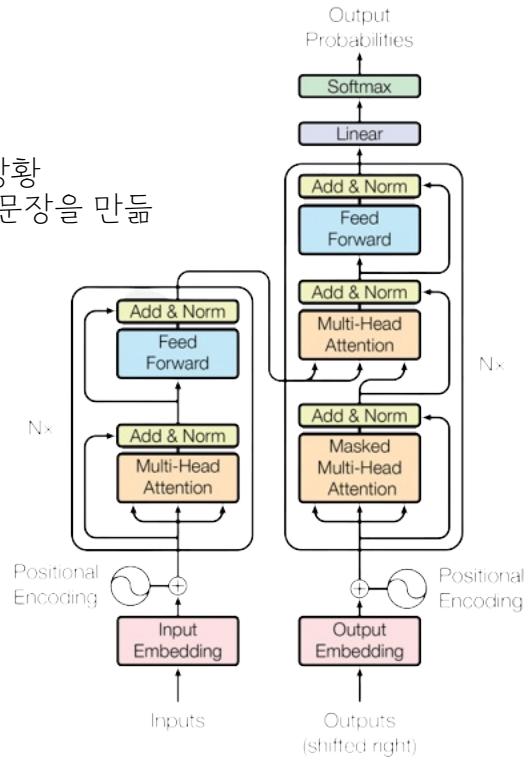
Generative Deep Learning

# 9 Transformers

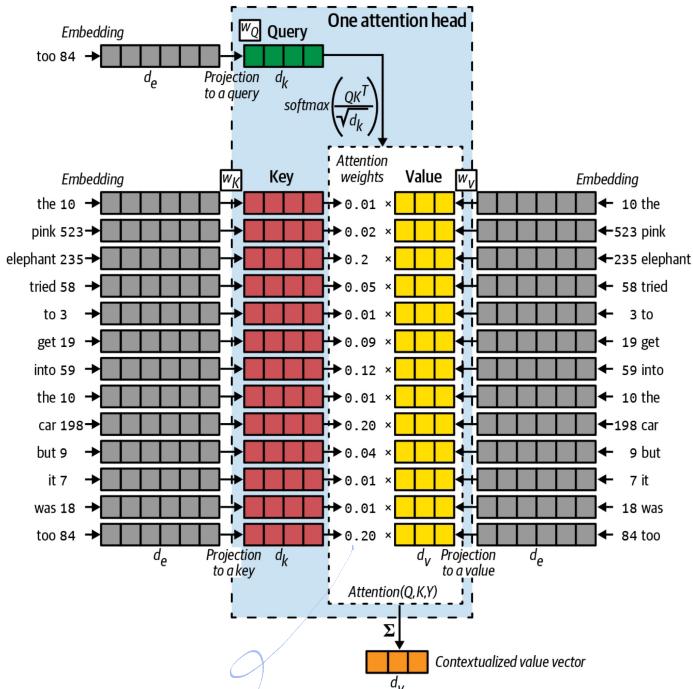
Yeseul Oh

# Transformer

- 순차모델링을 위해 어텐션 메커니즘에만 의존하는 신경망
- 병렬화가 용이하여 대규모 데이터셋에서 훈련 가능
- 번역하는 구조
  - \* Encoder: 번역당하는(source) 문장 - 전체정보 다 보고 있는 상황
  - \* Decoder: 번역할(target) 문장 - target 언어의 단어 순서대로 문장을 만듦



# Attention



기본 개념

ref > 같은 단어인자 . 단어의 유사도 X  
겹바나 관련이 있는지, 관련도 O

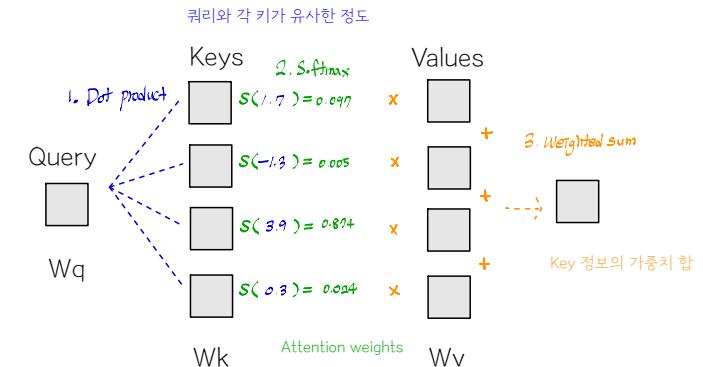


Figure 9-2. The mechanics of an attention head

?  
 $\neq 1$

〈비유〉 Google 검색 엔진  
검색어(Query)를 받아서 가진 웹사이트(Values)에서 검색어와 관련도(Weights from keys)를 구해 관련이 높은 정보들을 보여줌

Query → 

权重 1등  
Value 1

NVIDIA Blog Korea  
<https://blogs.nvidia.co.kr>  
트랜스포머 모델이란 무엇인가? (1) | NVIDIA Blog  
트랜스포머 모델은 문장 속 단어와 같은 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망입니다.  
어텐션(attention) 또는 ...

权重 2등  
Value 2

위키독스  
<https://wikidocs.net>  
16-01 트랜스포머(Transformer) - 딥 러닝을 이용한 자연어 처리 입문  
트랜스포머(Transformer)는 2017년 구글이 발표한 논문인 "Attention is all you need"에서 나온 모델로 기존의 seq2seq의 구조인 인코더-디코더를 떠르면서도, 논문의 ...

权重 3등  
Value 3

FFighting  
<https://ffighting.net>  
Transformer 논문 리뷰 - ChatGPT 모델의 근간 확실하게 이해하기  
2023. 9. 30. – Transformer 모델은 딥 러닝과 자연어 처리 분야에 큰 영향을 미쳤습니다. 이 모델은 별별 처리 능력과 위아닌 성능으로 다양한 NLP 작업에서 활용되고 ...

权重 4등  
Value 4

Wikipedia  
<https://en.wikipedia.org>  
Transformer (deep learning architecture)  
A transformer is a deep learning architecture developed by researchers at Google and based on the multi-head attention mechanism



# Multi-head Attention

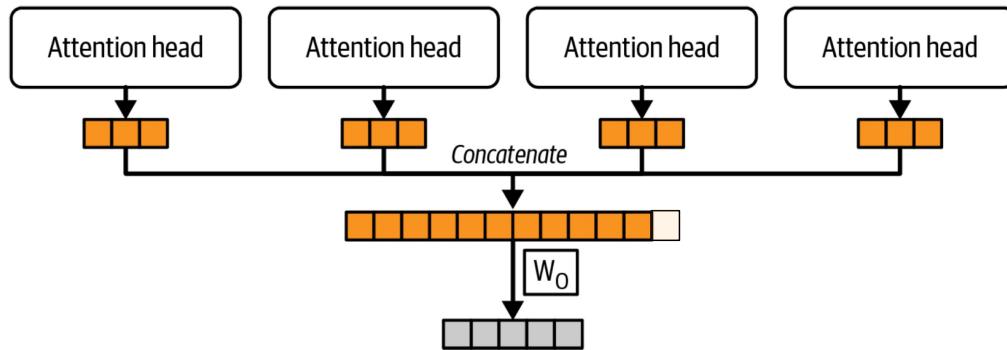


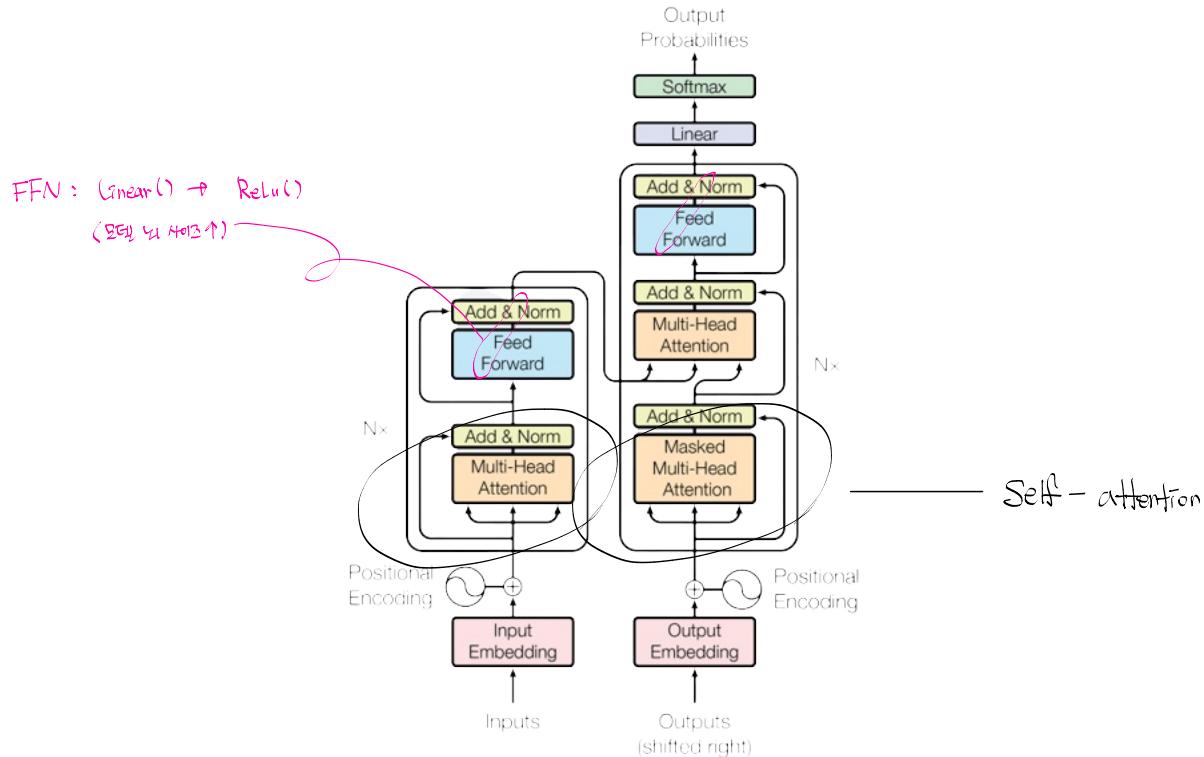
Figure 9-3. A multihead attention layer with four heads

Attention head마다 고유한 attention 메커니즘을 학습하여 총 전체가 더 복잡한 관계를 학습할 수 있다.

# 문제점

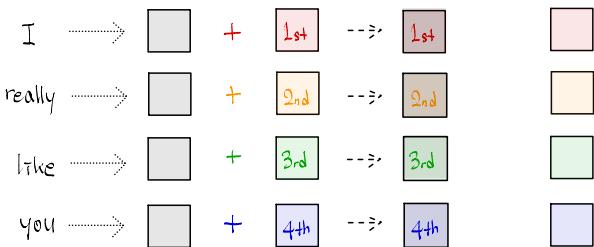
위 메커니즘은 query 와 key의 유사도를 각각 개별적으로 보기 때문에 단어들의 "순서"를 고려하지 못한다. → Positional encoding

위 메커니즘은 단어들의 순서는 고려할 수 있어도 "맥락" 파악이 불가하여 뜻이 다른 동음이의어는 구별 하지 못한다. → Self-attention

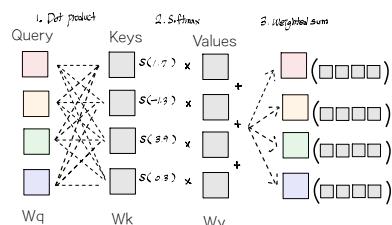
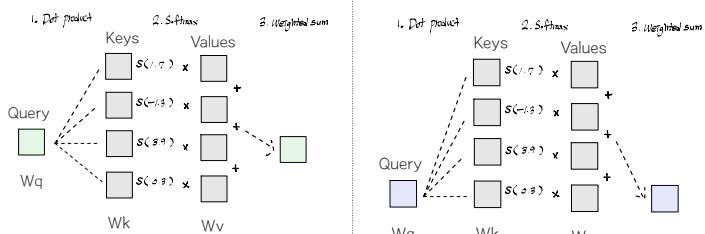
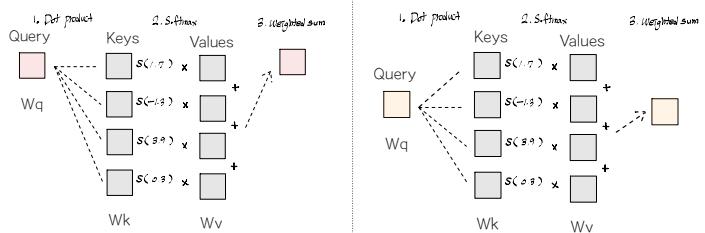


# Positional encoding

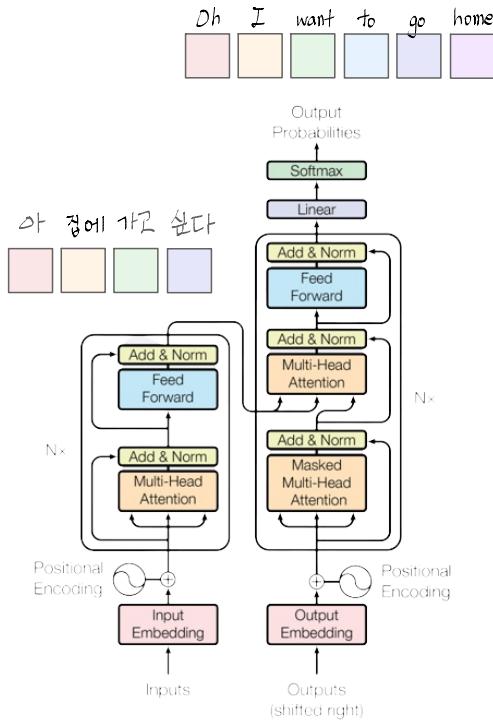
Token embedding  
단어나 토큰을 고정된 차원의 실수 벡터로 변환



# Self-attention



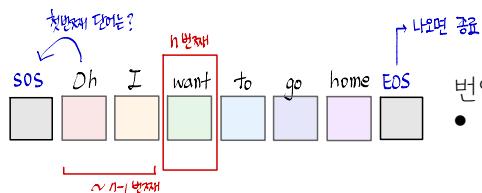
# 작동/학습 방식



아 집에 가고 싶다

아 집에 가고 싶다

- Encoder에 전부 넣음.
- Decoder의 중간 attention 의 key, value로 들어가서 한국어 문장 정보를 입력.

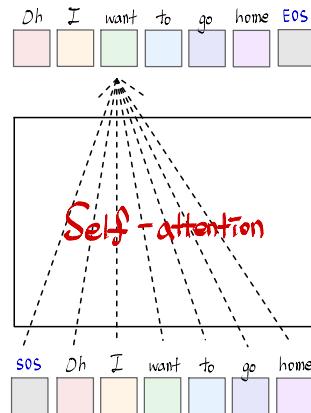
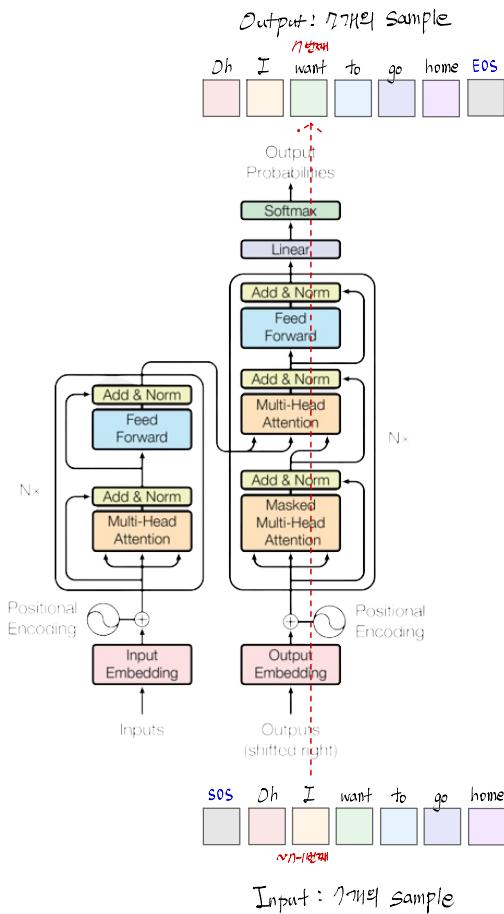


번역될 문장: 영어

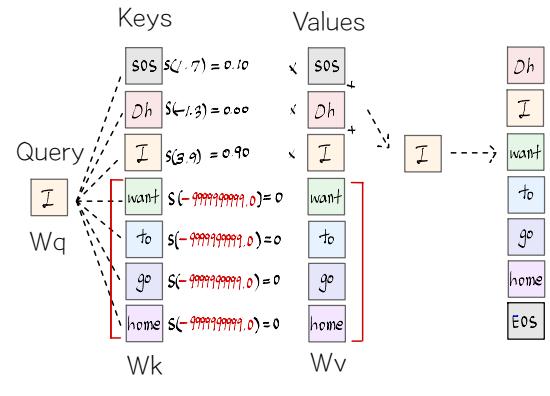
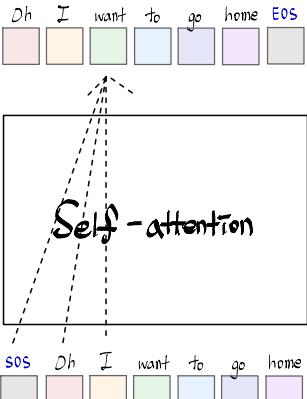
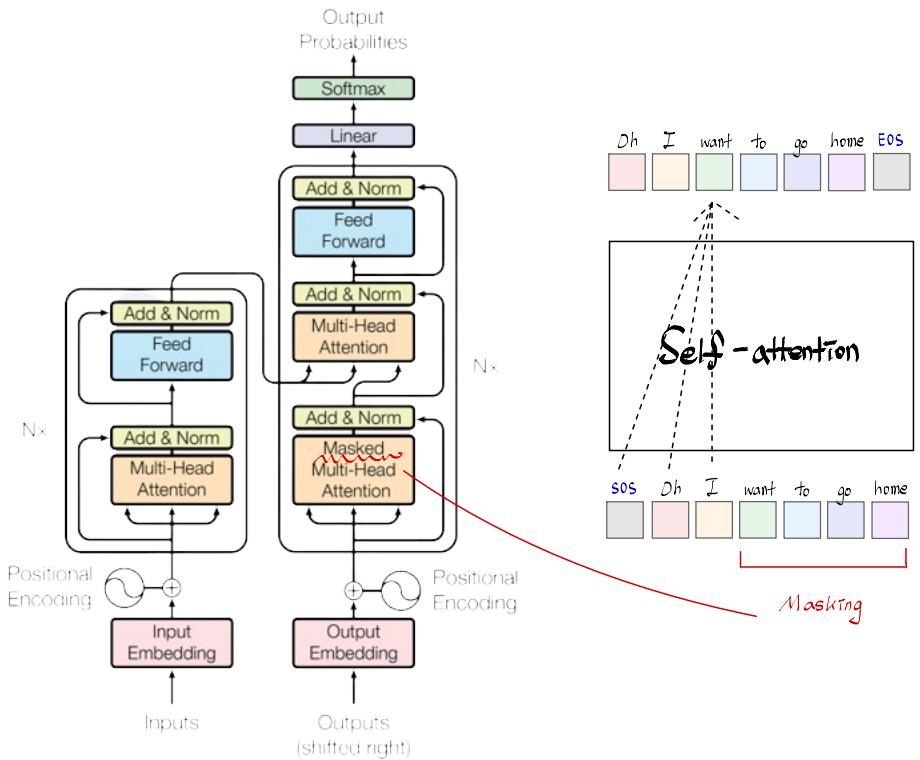
- Decoder에  $n-1$  번째 까지의 단어의 정보를 입력.  $n-1$  번째 단어 위치의 결과가  $n$  번째 단어를 예측하도록 함.
- Start/End of Sequence

학습 시:  $\sim n - 1$  개 단어 넣어서  $\rightarrow n$  번째 단어(정답) 예측  
병렬 학습 가능: 단어 개수만큼 동시에 주기

# 병렬 처리



# Masking



Masking: 절대값이 큰 음수로 교체

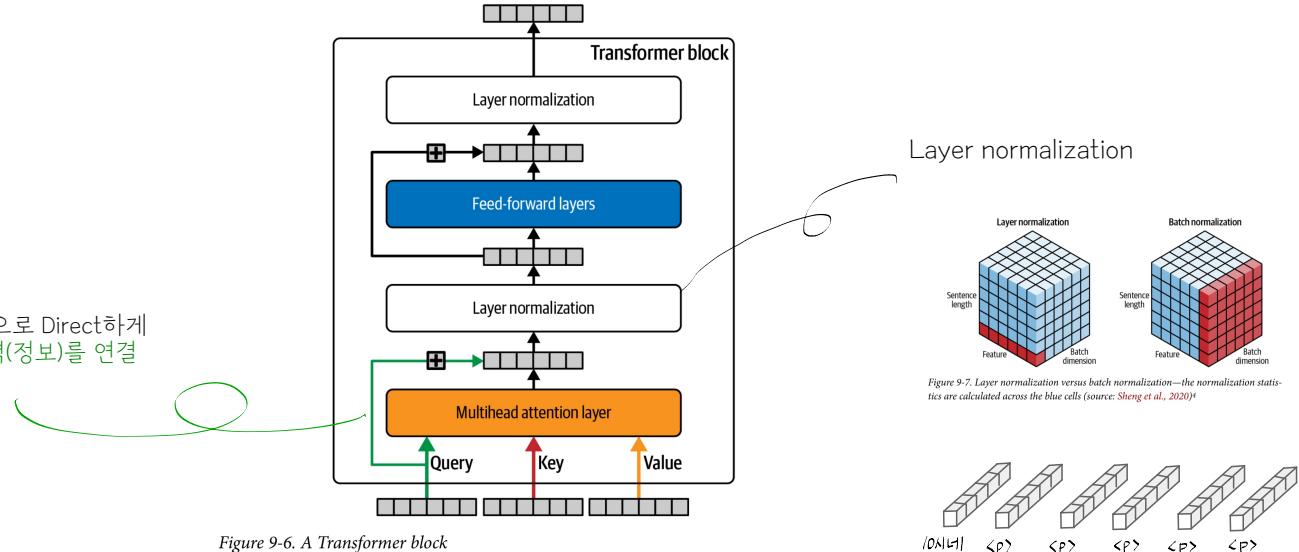
	Query												
Key	the	pink	elephant	tried	to	get	into	the	car	but	it	was	too
the	1.00	0.67	0.22	0.35	0.24	0.22	0.01	0.14	0.10	0.01			
pink	0.33	0.44	0.01	0.17	0.26	0.12	0.01	0.13	0.09	0.04	0.02		
elephant	0.34	0.15	0.27	0.20	0.05	0.20	0.15	0.04	0.05	0.11	0.23		
tried		0.65	0.33	0.11	0.20	0.16	0.05	0.07	0.17	0.05	0.05		
to			0.00	0.05	0.06	0.08	0.11	0.01	0.07	0.07	0.01		
get				0.36	0.22	0.01	0.07	0.05	0.16	0.05	0.09		
into					0.15	0.08	0.17	0.11	0.02	0.12			
the						0.07	0.11	0.03	0.01	0.08	0.01		
car							0.20	0.01	0.03	0.02	0.20		
but								0.07	0.17	0.00	0.04		
it									0.02	0.02	0.01		
was										0.24	0.01		
too											0.2		

Figure 9-4. Matrix calculation of the attention scores for a batch of input queries, using a causal attention mask to hide keys that are not available to the query (because they come later in the sentence)

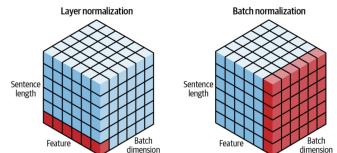
# Transformer block

## Skip connection

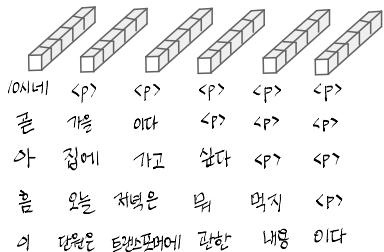
이전 Layer의 정보를 직접적으로 Direct하게 이용하기 위해 이전 층의 입력(정보)를 연결



Layer normalization



*Figure 9-7. Layer normalization versus batch normalization—the normalization statistics are calculated across the blue cells (source: Sheng et al., 2020)*



# Other transformers

- BERT
  - \* encoder transformer
  - \* 문장에서 누락된 단어의 전후 문맥을 고려하여 누락 단어를 예측
- GPT
  - \* decoder transformer
  - \* 한 번에 하나의 토큰씩 텍스트 문자열을 생성
- T5
  - \* encoder-decoder transformer
  - \* text to text framework번역, 언어의 적합성, 문장 유사성, 문서 요약

