



Longformer

# SLIDING WINDOW ATTENTION(SWA)

Longformer: The Long-Document Transformer

신혜빈

2024.01.09.

# CONTENTS

목차

1

**Longformer**  
등장 배경

2

**Longformer's  
attention  
mechanism**

3

**Attention  
Pattern**

4

**정리**



# Longformer의 등장 배경



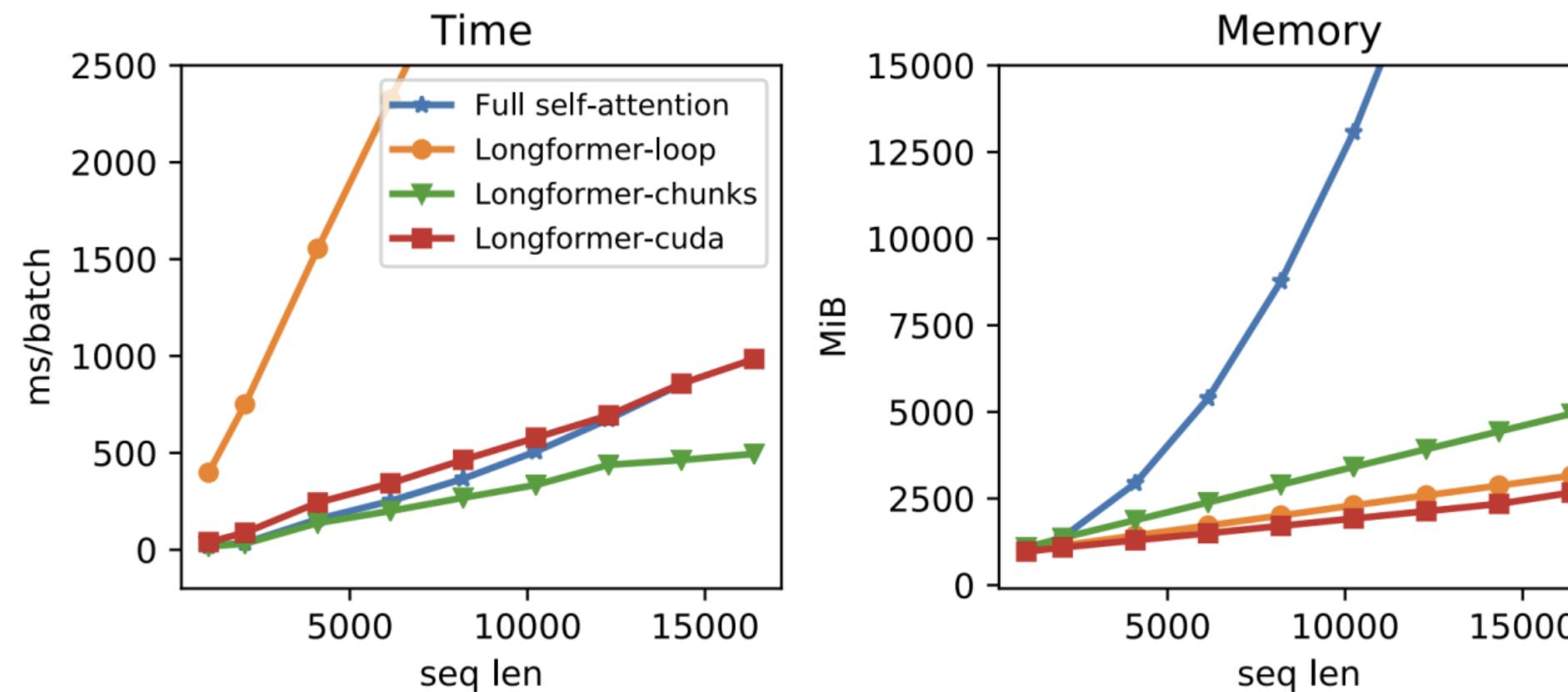
## 긴 문서에 대한 효율적 처리 필요

기존 트랜스포머 기반 모델들의 **셀프 어텐션 연산**은  
시퀀스 길이에 따른 계산, 메모리가 **제곱**으로 증가





# Longformer의 등장 배경



시퀀스 길이에 "선형적으로"  
확장되는 어텐션 메커니즘 도입





# Longformer의 등장 배경

Long document classification,  
Question answering(QA)



등에 유리

시퀀스 길이에 "선형적으로"

확장되는 어텐션 메커니즘 도입





# Longformer's attention mechanism

## Local windowed attention

주로 문맥 표현을 구축하는 데 사용

+

## Global attention

예측을 위한 전체 시퀀스 표현을 구축하는 데 사용





# Longformer's attention mechanism

## Windowed local-context self-attention

모델이 입력 시퀀스 내의 각 토큰이  
주변의 특정 범위 내 토큰들과만 상호작용하도록 제한

+

## Global attention

모델이 전체 입력 시퀀스에 걸쳐  
중요한 토큰들에 주목하게 함





# Time and memory complexity

**The original Transformer model**

$$O(n^2)$$



\*n: input sequence length



# Time and memory complexity

## The original Transformer model

수천 개 이상의 토큰을 포함하는  
문서 처리에서 어려움을 겪음



\*n: input sequence length

# KEY POINT

Longformer



## Sparsify

일부 "중요한"  
토큰 쌍에만  
어텐션 적용



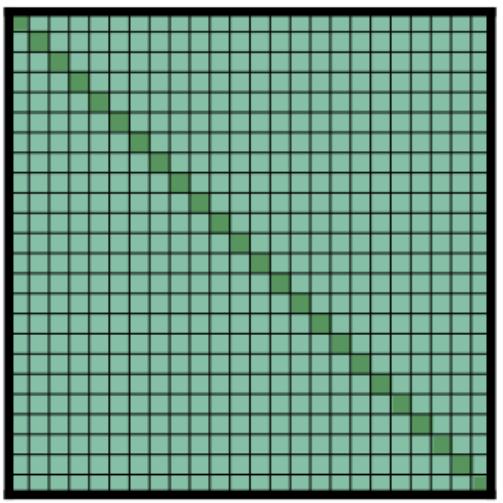
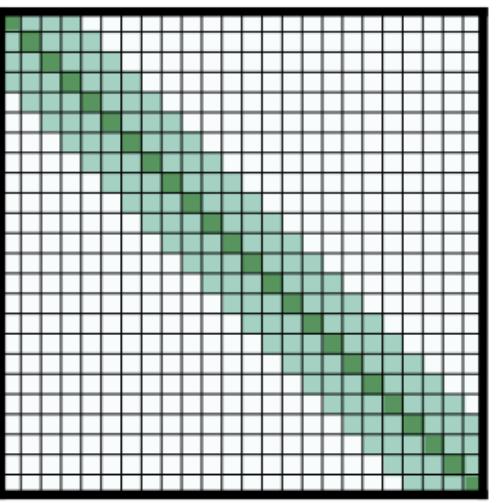
## Attention Pattern

입력 시퀀스와  
"선형적"으로 확장되어  
긴 시퀀스에 대해 효율적

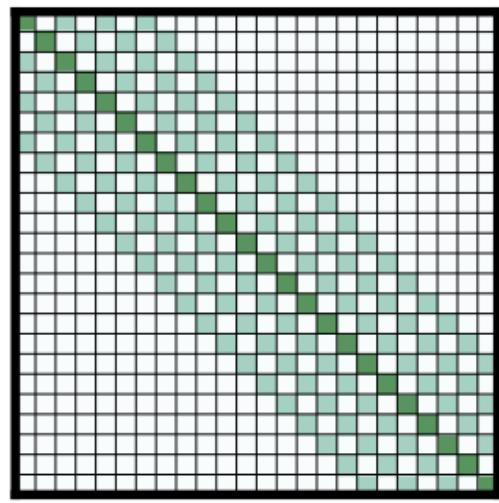




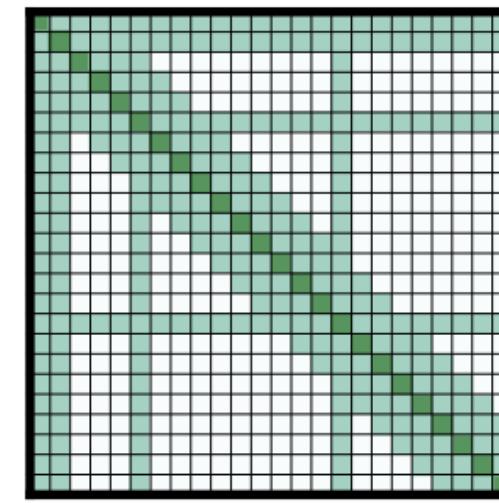
# Attention pattern

(a) Full  $n^2$  attention

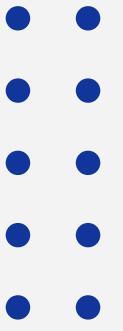
(b) Sliding window attention



(c) Dilated sliding window

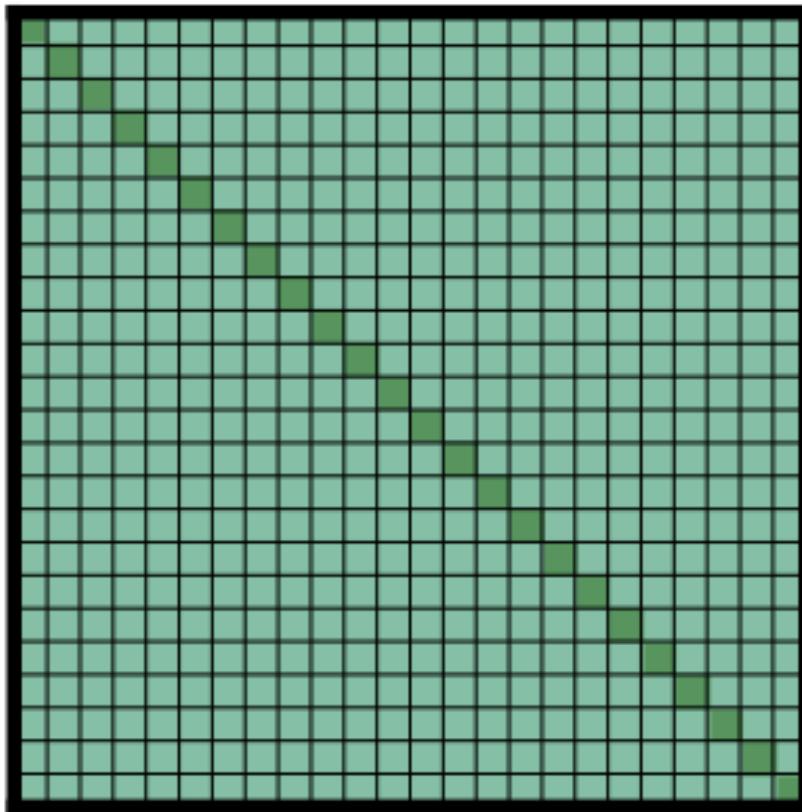


(d) Global+sliding window



# Attention pattern

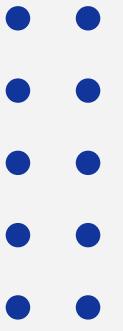
## Full self-attention pattern



(a) Full  $n^2$  attention

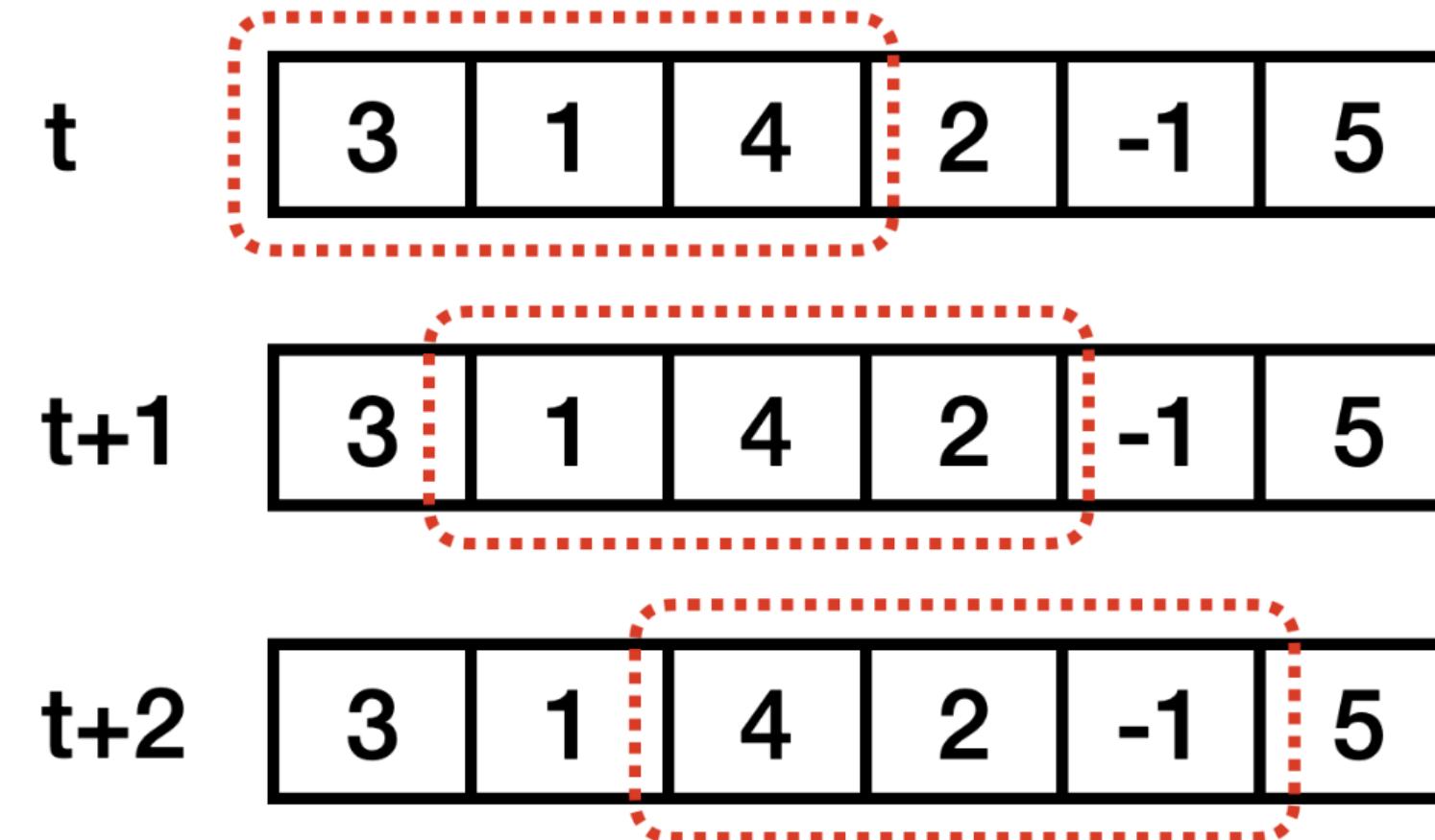
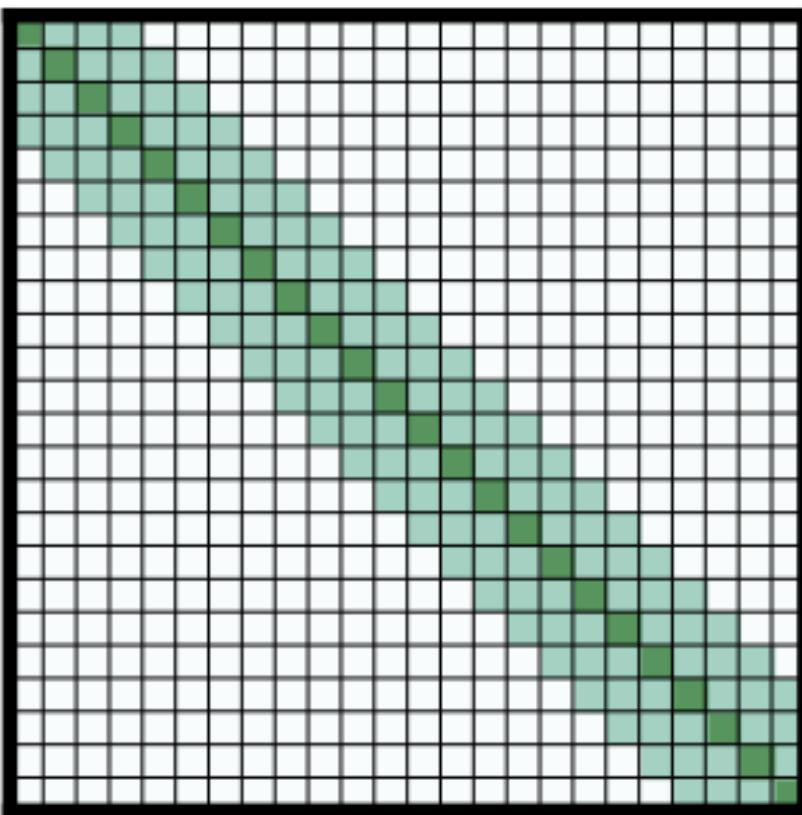
- 모든 토큰 쌍 간의 어텐션 계산
- computation complexity  
:  $O(n^2)$

\*n: input sequence length



# Attention pattern

## Sliding window attention

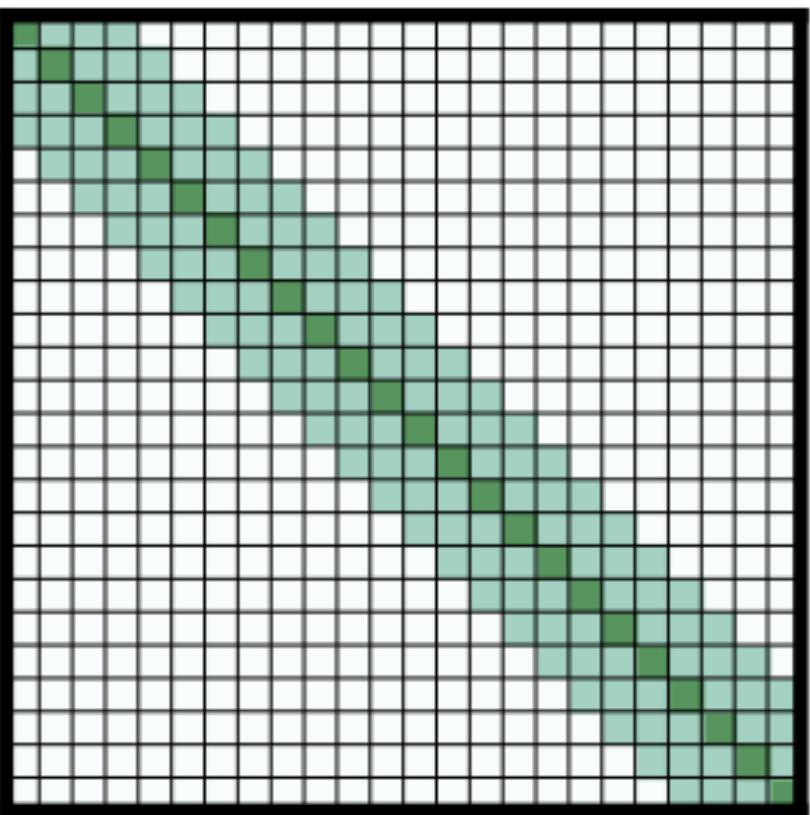


(b) Sliding window attention



# Attention pattern

## Sliding window attention



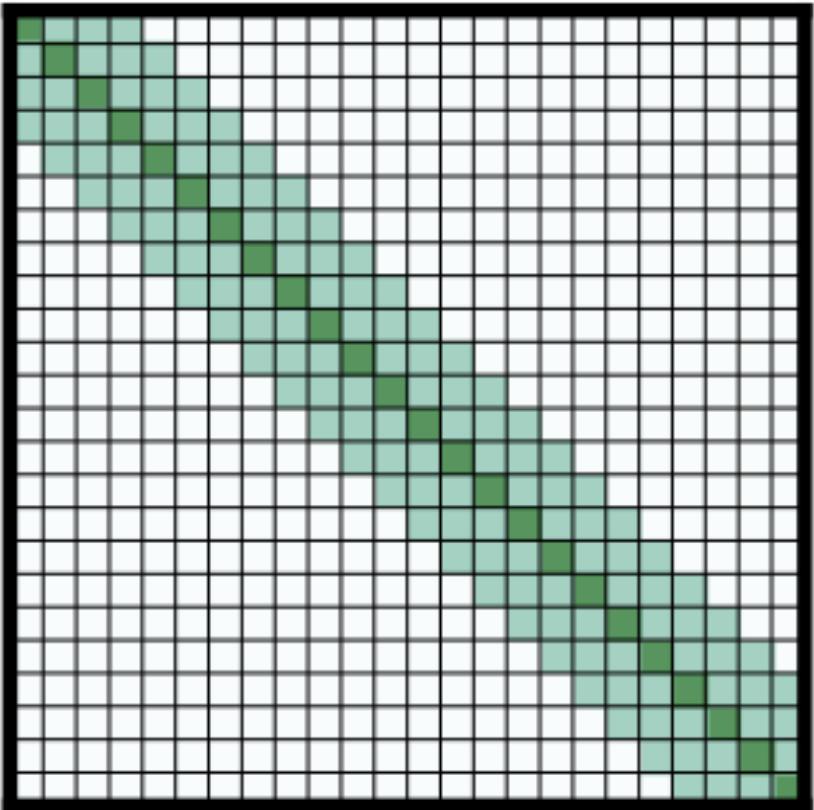
(b) Sliding window attention

- 고정된 window size ' $w$ '가 주어졌을 때,  
각 토큰들은 양옆에  $1/2w$  토큰에 주의를 기울임
- window attention을 여러 층에  
쌓아 올림으로써 큰 수용 영역 형성
- 최상위 층은 모든 입력 위치에 접근 가능
- 전체 입력에 걸친 정보를 통합한 표현을  
구축할 수 있는 능력을 갖추게 됨



# Attention pattern

## Sliding Window Attention



- computation complexity  
:  $O(n \times w)$
- the receptive field size at the top layer  
:  $I \times W$

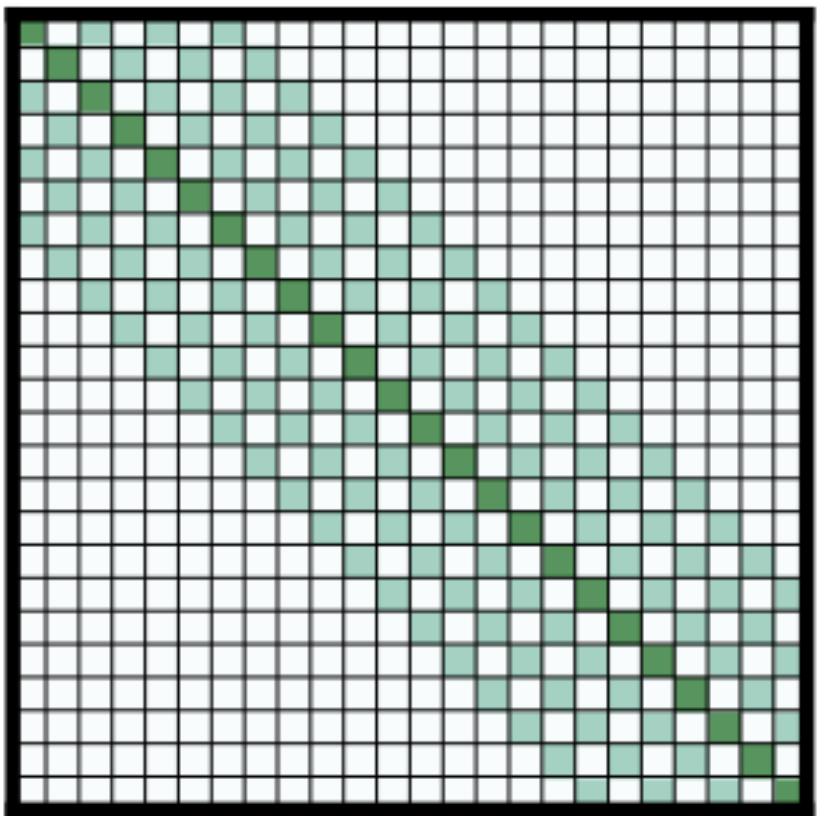
(b) Sliding window attention

\*응용 프로그램의 요구에 따라 각 층에서 window size  $w$ 를 다르게 설정 가능



# Attention pattern

## Dilated Sliding Window Attention



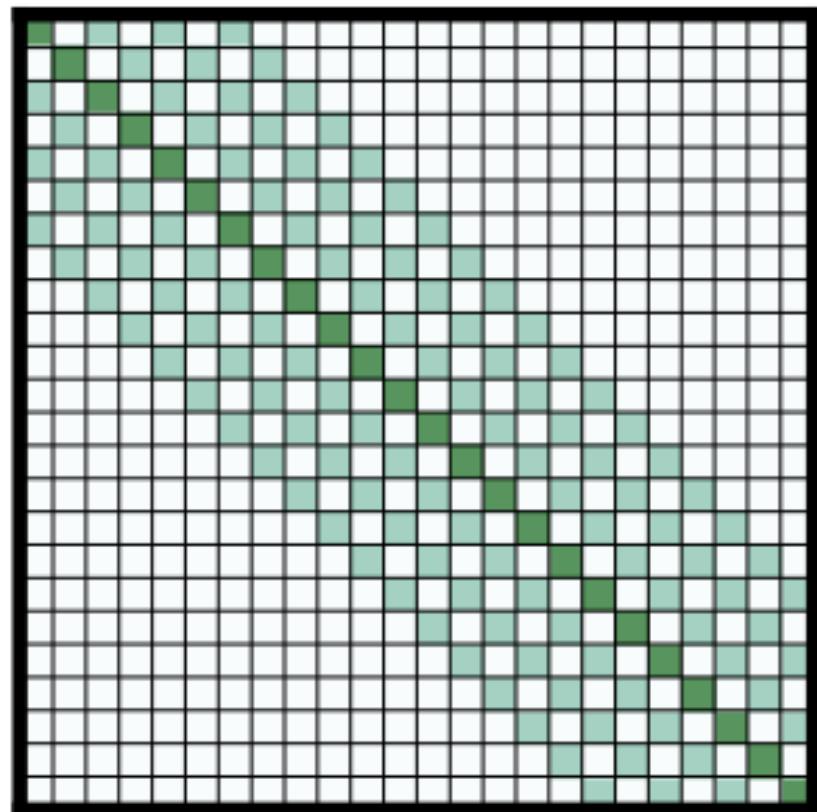
(c) Dilated sliding window

- receptive field는 늘리면서 계산량은 증가시키지 않게 하기 위한 확장(Dilated)
- gaps of size dilation ' $d$ '
- multi-head attention에서, 각각 헤드에 대해 다른 dilated attention score를 적용해 계산 가능
- 일부 헤드는 dilation없이 가까운 토큰에만 집중, dilation이 있는 헤드는 긴 컨텍스트에 집중



# Attention pattern

## Dilated Sliding Window Attention



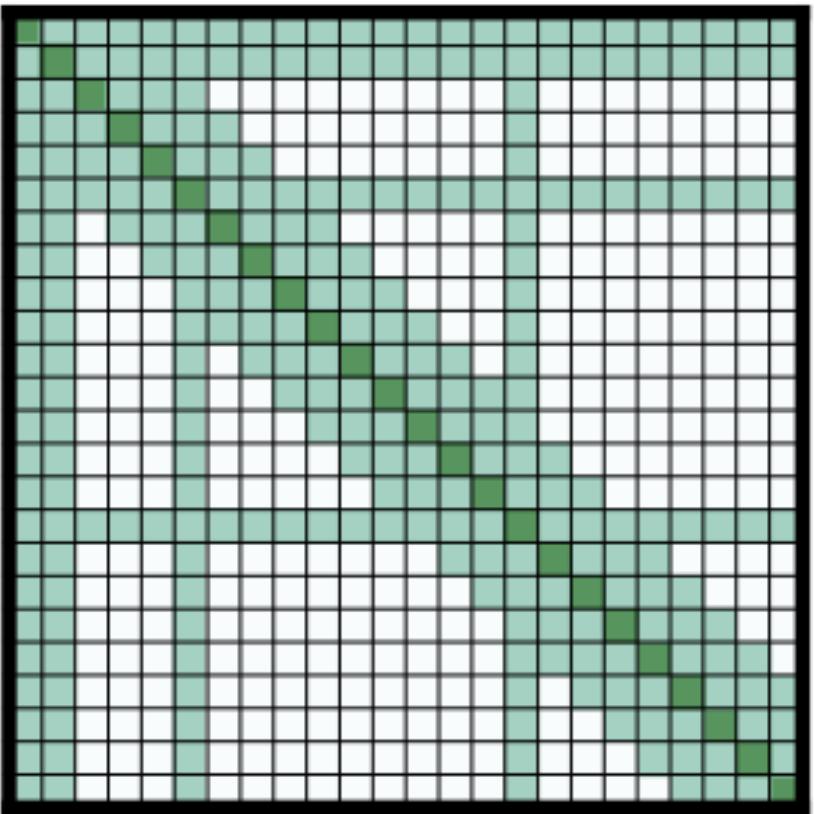
- the receptive field size  
 $: l \times d \times w$

(c) Dilated sliding window



# Attention pattern

## Global Attention



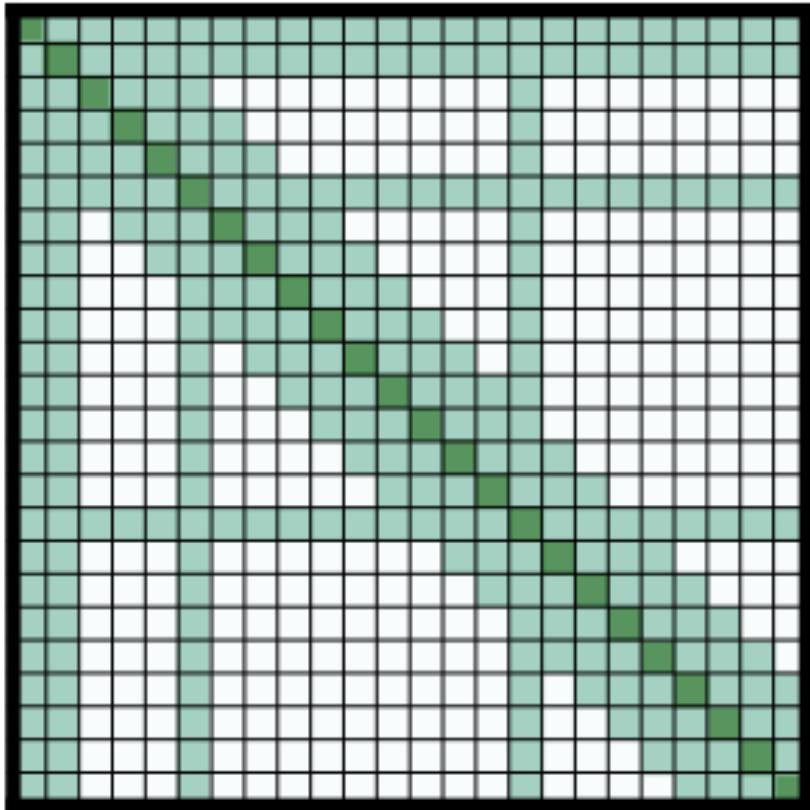
- Sliding window attention과  
Dilated sliding window attention의 한계점  
: local context만 바라봄
- Global Attention의 보완  
: 긴 텍스트에 대한 task 효율적 처리

(d) Global+sliding window



# Attention pattern

## Global Attention



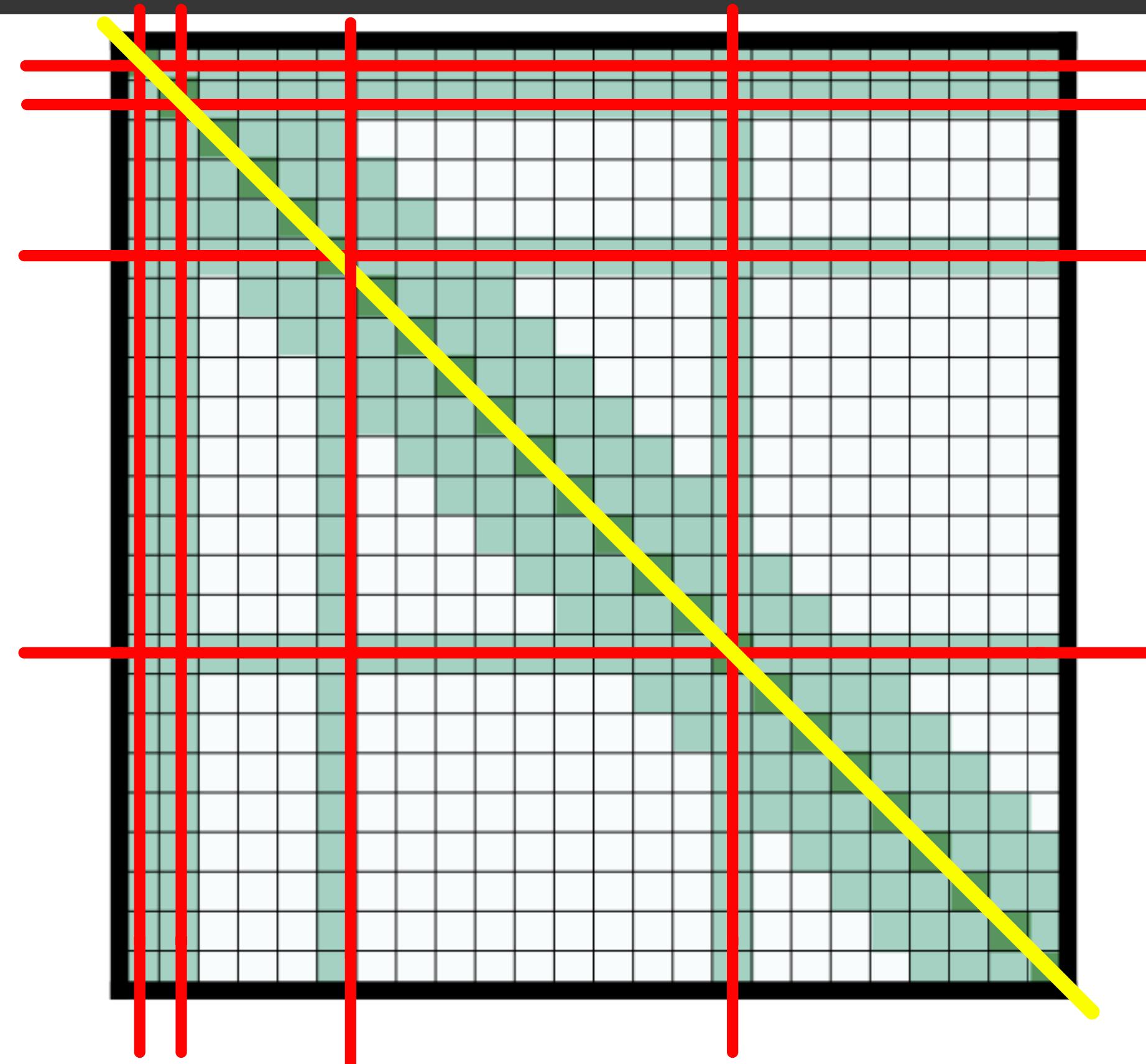
(d) Global+sliding window

- BERT
  - Classification: [CLS] 토큰에 global attention
  - QA: 모든 질문 토큰에 global attention
- Global + sliding window 형태
  - Global: 전체 시퀀스에 걸쳐  
다른 모든 토큰에 주의를 기울임
  - sliding window: 각 토큰의 로컬 컨텍스트를 포착

## Global attention



(d) Global+



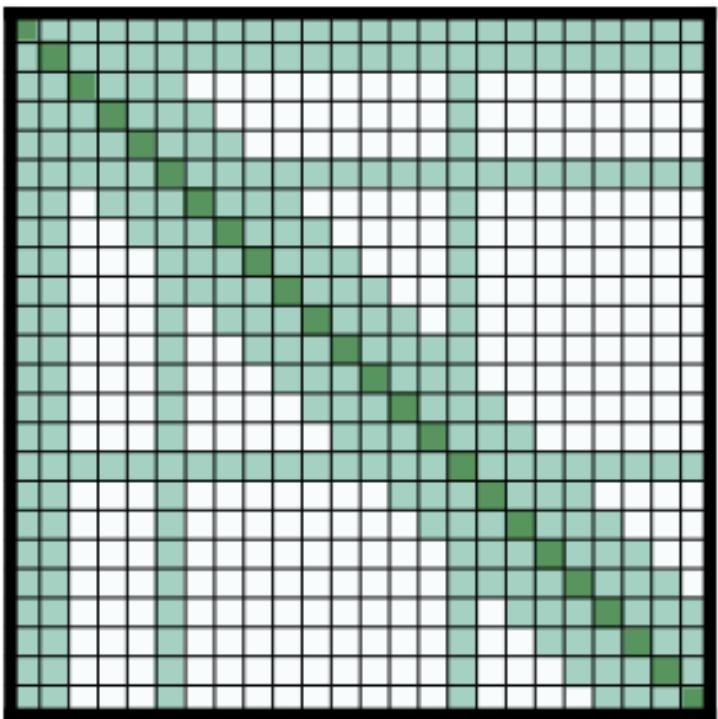
(d) Global+sliding window

Sliding  
window  
attention  
pattern



# Attention pattern

## Global Attention

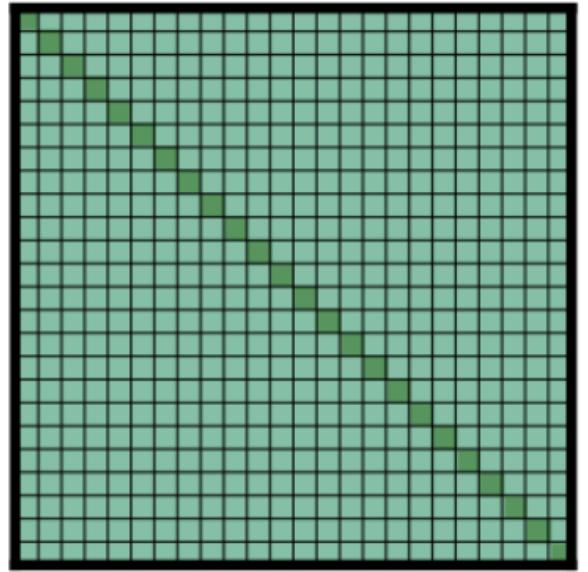


(d) Global+sliding window

- computation complexity  
:  $O(n)$
- 부연 설명
  - Global attention  
: 시퀀스 내 소수의 중요한 위치에만 적용
  - Local attention  
: 각 토큰의 주변에만 적용
- 전체 시퀀스에 대한  
모든 토큰 쌍 간의 어텐션을 계산할 필요X

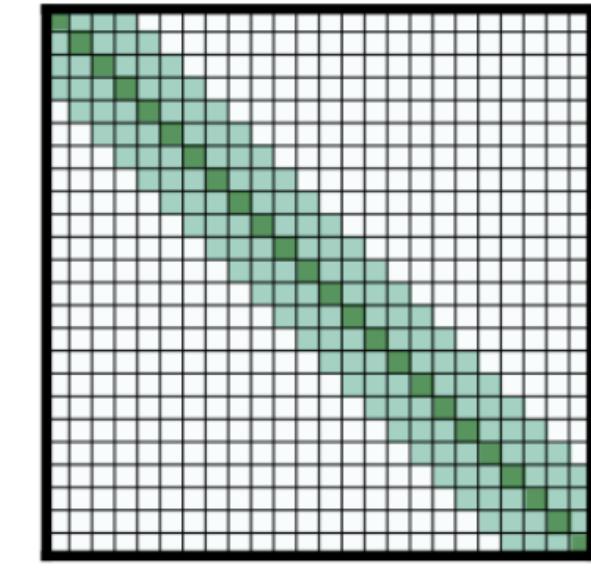


# 정리



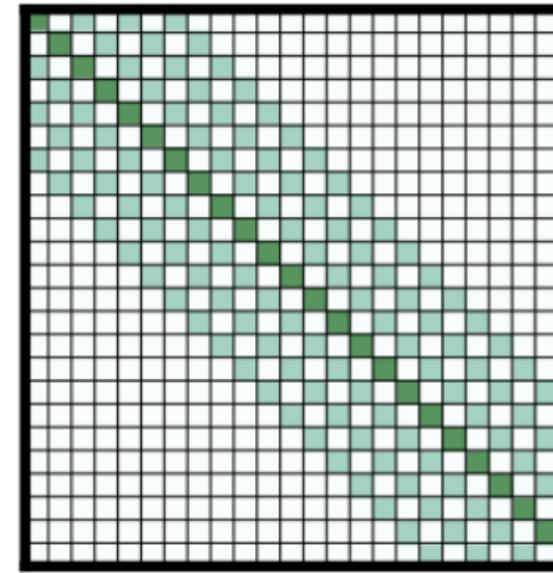
## Full $n^2$ attention

모든 토큰 쌍 간의  
어텐션을 계산함으로써  
입력 시퀀스 길이에 따라  
제곱되는 계산량과 메모리  
 $O(n^2)$



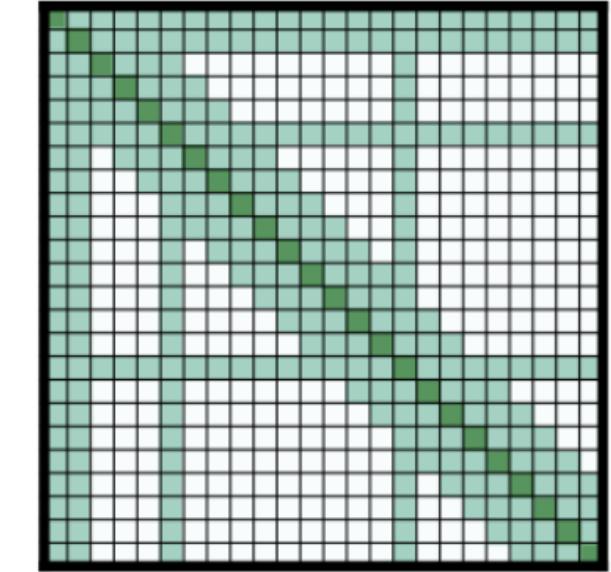
## Sliding window

특정 토큰 앞뒤로  
 $1/2w$ 만큼의 토큰에  
대해 어텐션 계산  
 $O(n \times w)$



## Dilated Sliding Window

모델이 입력 데이터의  
근접한 부분에 있는  
중요한 정보를 포착



## Global attention

긴 텍스트에 대한  
task 작업에 용이  
 $O(n)$



**MISTRALMODELSTUDY**



THANKS FOR WATCHING  
**E.O.D.**

