

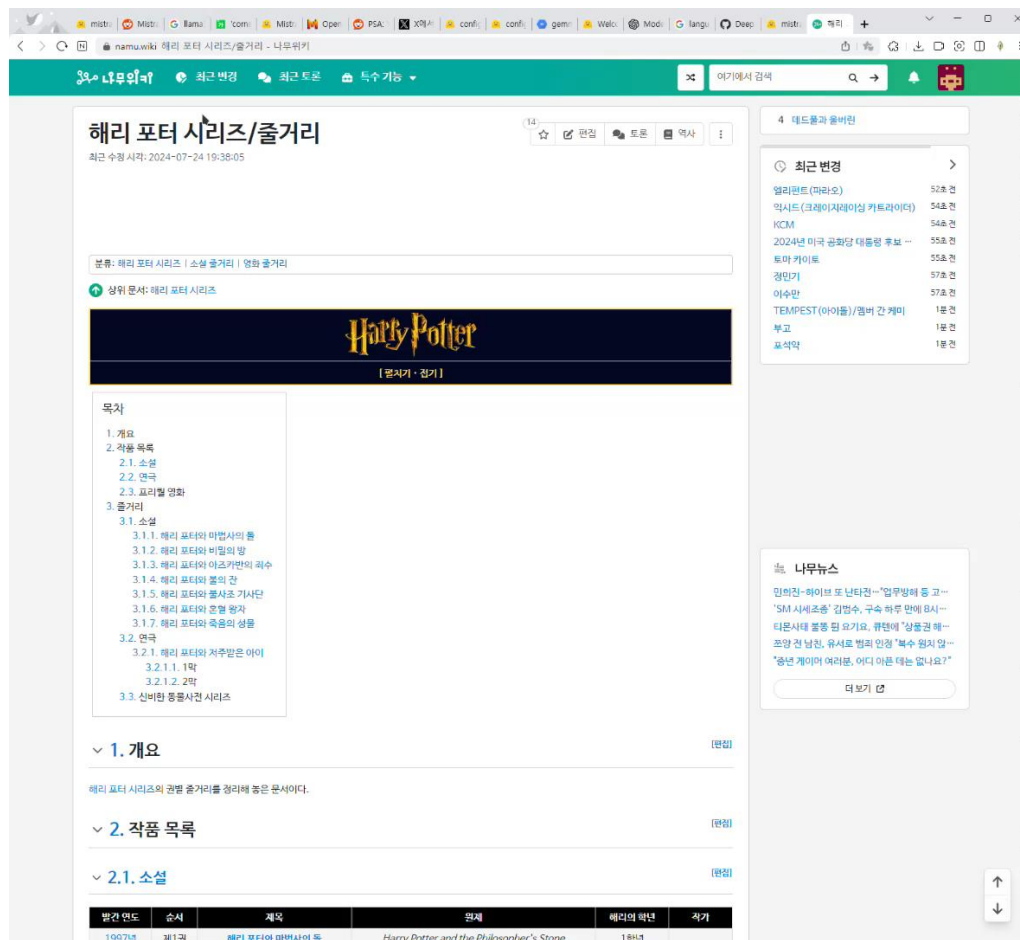
# How to deal with long sequence in self-attention

어떻게 긴 시퀀스가 셀프 어텐션에 들어갈 수 있을까?

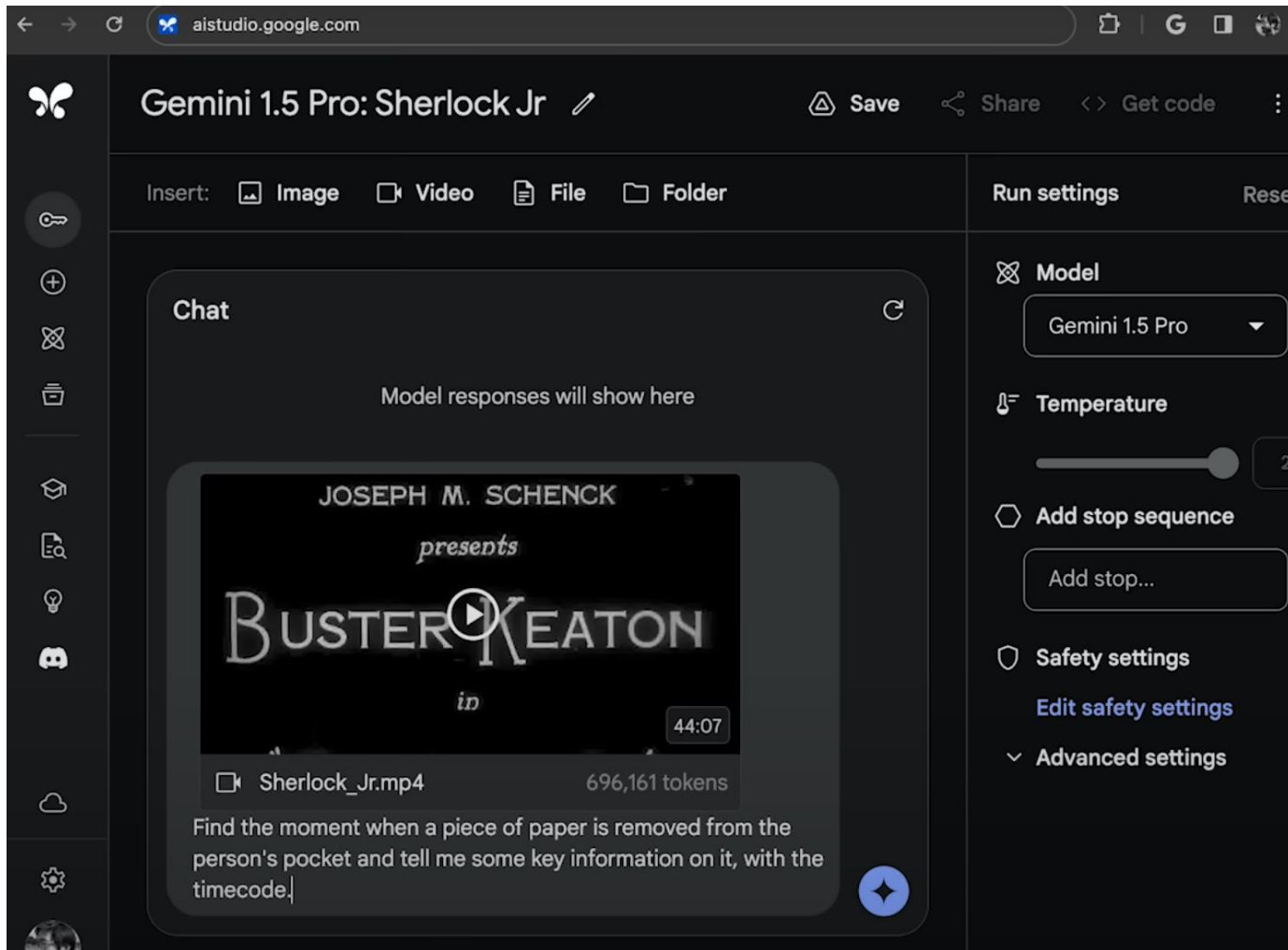
# 최신 모델 context window 동향

- Llama 3 → Llama 3.1 (8K → 128K)
- Mistral xxx → Mistral NeMo (32K → 128K)
- GPT-4-0613 → GPT-4-1106-preview (8K → 128K)
  
- Claude 2.1 (200K)
- Gemini 1.5 Pro (2M)
- Command R+ (128K)
  
- HyperClova X: 4K...

# 왜 context window가 크면 성능이 좋을까



# 사진, 비디오까지 고려한다면...?



44분짜리 영상에서

'어떤 사람의 주머니에서 종이  
가 떨어지는 순간을 찾아 묘사  
해라  
또 사건이 일어난 영상 시각은?'

이라고 묻고 있음

# 셀프 어텐션...

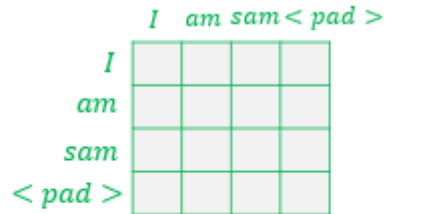
- $\mathbf{W}_Q$
- $\mathbf{W}_K$
- $\mathbf{W}_V$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_K}}\right)\mathbf{V}$$

# 긴 시퀀스 처리가 힘든 이유?

Context window: 4K

(그림은 4개지만 4K개라고 생각하고 넘어가자)

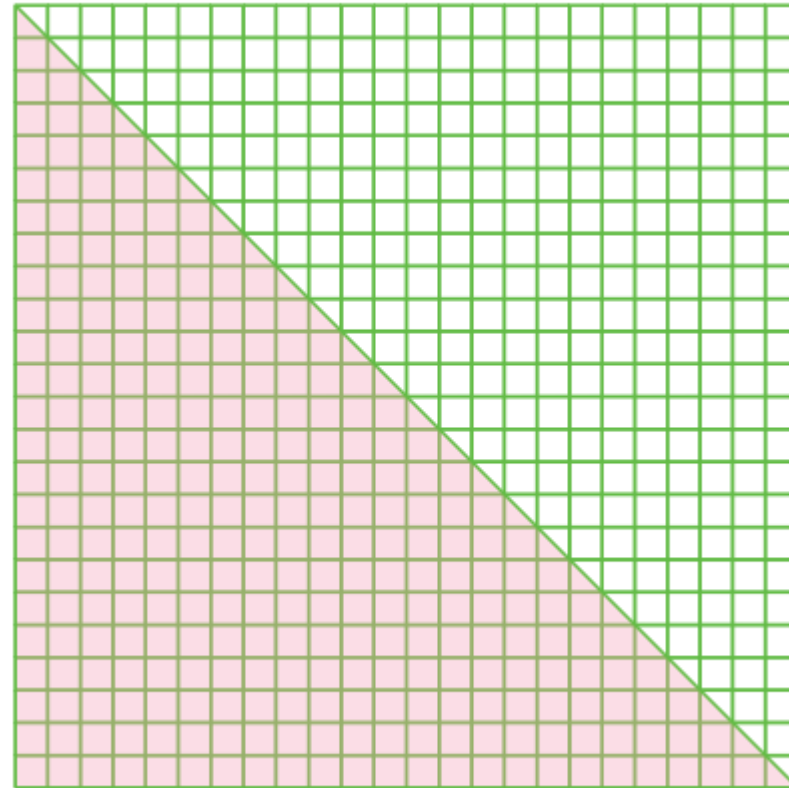


Attention Score Matrix

$$4K * 4K = 16M$$



Context window: 128K



Attention Score Matrix

$$128K * 128K = 16.4G$$



“셀프 어텐션 시 메모리, 연산량은 context window의 제곱에 비례”

# 용어 정리: context window란?

## “모델이 훈련할 때 사용한 시퀀스 길이”

셋은 콜라주가 대해 알아보려 여러 조사를 해 보나 나오지 않는다. 그 와중에 크리스마스가 찾아오고,<sup>[29]</sup> 해리는 크리스마스 선물로 헤그리드가 보낸 특이한 나무 리프도, 더덕리 가족이 보낸 50원스 동전, 위즐리 부인이 보낸 달걀 피자와 스위트, 헤르미온드가 보낸 개구리 초콜릿을 열어 보고 마지막 선물을 뽀는 데 발견한 이라는 그 소도 안에서 어서히 찾아낸 부엉이 알도를 선물받고, 이를 이용해 콜라주에 대해 찾아보려 도서관에 갔다 돌아오는 길에 헤매다가 한 방 안에 있는 거울을 보게 된다. 울림에도 그 거울을 보자 해리의 뒤에 부모님을 비롯한 가족들을 보고, 다음날 혼과 함께 가지만, 혼은 그 회랑 테지를 달고 기숙사 우승컵과 허디스 우승컵을 든 자신을 보게 된다. 그 거울에 빠져버린 해리는 또 찾아가는데, 이번엔 그 방에서 덩블도어 교수들 만나게 되고, 그 거울이 진실의 자기 모습이 아니라 마음 속의 모습을 보여 주는 '소망의 거울'이라는 것을 알게 된다.<sup>[30]</sup> 그리고 다시는 거울을 찾지 않았다고 약속한다.<sup>[31]</sup> 나가기 전, 해리는 덩블도어에게 교수님은 이 거울을 보면 무엇이 보이냐고 물었고 덩블도어는 자신이 이 거울을 보면 정말 한 할례를 물고 있는 자신의 모습을 본다고 대답한다.<sup>[32]</sup>

여길 후, 네빌이 말포이와 달리 유기 주위에 당한 채 그리핀도르 후계설까지 풍문 휘어온다. 그런 네빌을 위로해 주기 위해 헤르미온드가 크리스마스 선물로 준 것 중 마지막 개구리 초콜릿을 뜯은 해리는 금탕열차 데저팅 덩블도어를 볼게 되는데, 그 뒷면의 내용에서 니콜라 플라멩가를 발견하고, 그가 아임사의 통을 만드는 유일한 연금술사이며, 665살이라는 걸 알게 된다. 헤그리드가 713년 금고에서 꺼낸 것, 그린고츠의 도둑이 훔쳐간 것, 복숭아가 지키고 있는 것은 불로장생의 약, 마법사의 용이었던 것이다.

2번째 라디지 경기, 우물부르와의 경기인데, 실은 온 스페이프로, 덩블도어 교수가 관전하러 나왔다. 그리핀도르는 스페이프로 편마 관성에도 불구하고 해리와 영활만으로 승리했다. 경기 후 해리는 아무도 몰라

### Context Window

시점 기간, 셋은 헤그리드가 리커 울드윈에서 만난 시종사자 카르타임으로 용의 용을 얻어냈다 는 걸 알게 된다.<sup>[33]</sup> 용의 알은 온 주된하고, 헤그리드는 그 태어난 새끼용의 이름을 노브르로 짓는다.<sup>[34]</sup> 그러나 헤그리드의 집에서 용을 기르는 것은 불가능할 뿐더러<sup>[35]</sup> 용이 노브르에게 손가락을 물리고 중독되어 입원을 한 대다가<sup>[36]</sup> 말포이가 용을 피버렸기 때문에 노브르를 웨이너리에서 용을 연구하는 용의 왕 윌리엄에게 넘겨주기로 한다. 그러나 말포이가 보낸 편지를 끼워 둔 용의 책을 말포이가 가져가버리는 바람에 그가 계획을 알아버린다. 해리 말포이가 계획은 말포이가 계획은 말포이도 투명 말도의 존재는 모른다는 점을 노려 아방에 투명 말도를 이용해 천문함으로 단견히 용을 배달하였다. 그러나 나무 산란 나이지 투명 말도를 두고 오는 바람에 울금 시간에 기숙사를 나간 것이 알려지고, 용에 대해 고자살하기 위해 역시나 밖에 나온 말포이와 함께 해리, 헤르미온드, 그리고 말포이가 나왔다고 주의를 주러 나온 네빌까지 각각 기숙사 점수 50점씩 도합 150점을 감점당하고 징계도 받게 된다.

장기는 헤그리드의 함께 금지된 숲에서 그의 말을 듣는 것이었다. 최후로 유니콘이 자주 죽는데 그 이유를 알아내는 것이 임무였다.<sup>[37]</sup> 처음에는 헤그리드, 해리, 헤르미온드가 한츠, 네빌, 말포이, 그리고 헤그리드의 개 같이 한츠로 움직였다. 그 와중에 금지된 숲에 천타우로스도 만나 잠시 이야기도 하는데, 이때 네빌 쪽에서 이상을 의미하는 불꽃을 쏘아 올린고 헤그리드가 바로 달려가본다. 하지만 그건 말포이가 네빌을 놀리키는 장난을 치서 올린 네빌이 쏘아 올린거였고, 헤그리드는 네빌을 자신 의 조에 넣고 해리를 말포이와 가깝다. 말포이와 이동을 하다 유니콘의 피타극을 보고, 그걸 따라 움직이던 해리는 어떤 무기를 쓴 왕실이 유니콘의 피를 빨아먹는 것들 보는데<sup>[38]</sup>, 동시에 이마의 흉터에 엄청난 고통을 받는다. 다들히 천타우로스 피탄체가 해리를 구해준 뒤 헤그리드에게 데려가 준다. 해리는 그린고츠의 도둑과 트롤을 물리는 시간의 배후에 블드보트가 있다고 생각한다. 그리고 진실로 돌아가자, '만살을 대비하여'라는 해지와 함께 다시 투명 말도가 돌아와 있다.

그렇게 의욕에 빠진 상태에서 기말고사를 치르는 와중, 해리의 용더는 계속 아프고, 헤그리드에게 찾아가 용의 알을 온 사람에게 대해 이야기하는데<sup>[39]</sup> 알고 보니 그 사람은 처음부터 복숭아에게 지대한 관심을 갖고 있었고 결국 헤그리드가 숨기운다. 그자에게 복숭아를 치내하는 방법<sup>[40]</sup>을 알려줬었다. 이를 알게 된 해리와 혼, 헤르미온드는 그 사람이 스페이프로 블드보르라고 확실하게 아임사의 용을 용치려 할 것이라는 걸 덩블도어 교수에게 알려주지만, 그는 어떤 경우의 효율을 받고 감시자를 할 한 상태였다. 대신 맥그나일 교수에게 알려주지만 그녀는 그것은 안전하며, 또 다시 증거명확하면 또 기숙사 점수 50점을 감점하겠다고 한다.

해리, 혼, 헤르미온드는 이방곳곳과 투명 말도를 쓰고 아임사의 통을 찾아온 거리고 하나, 네빌이 더 이상의 감정은 안 된다는 상충사를 기록한다. 헤르미온드가 급세 풍작그만 주문으로 네빌을 제압하고 서둘러 복숭아의 방으로 향한다. 헤그리드가 크리스마스 선물로 지던 피리프 음악을 들려주자 복숭아가 잠들었고 지하실 문으로 떨어진 자카의 뒷이만 석물에 흠이 단단히 묶이는데, 헤르미온드가 통을 열어 탈출한다.

# 용어 정리: context window란?

## “모델이 훈련할 때 사용한 시퀀스 길이”

셋은 콜라주에 대해 알아보려 여러 조사를 해 보나 나오지 않는다. 그 외중에 크리스마스가 찾아오고,<sup>[29]</sup> 해리는 크리스마스 선물로 헤그리드가 보낸 특이한 나무 리프도, 더덕리 가족이 보낸 50원스 동전, 위즐리 부인이 보낸 달걀 피자 와 스퀘터, 헤르미온드가 보낸 개구리 초콜릿을 열어 보고 마지막 선물을 뜯는데 발견한 이라는 그 소도 안에서 어서히 것이라는 투명 알도를 선물받고, 이를 이용해 콜라주에 대해 찾아보려 도서관에 갔다 돌아오는 길에 헤매다가 한 방 안에 있는 거울을 보게 된다. 울 말기도 그 거울을 보자 해리의 뒤에 부모님을 비롯한 가족들을 보고, 다음날 종과 함께 가지만, 혼 혼고 회랑 테지를 달고 기숙사 우승컵과 허디지 우승컵을 든 자신을 보게 된다. 그 거울에 비쳐버린 해리는 또 찾아가는데, 이방엔 그 방에서 알등도어 교수들 만나게 되고, 그 거울이 진실의 자기 모습이 아니라 마음 속의 소망을 보여 주는 '소망의 거울'이라는 것을 알게 된다.<sup>[30]</sup> 그리고 다시는 거울을 찾지 않겠다고 약속한다.<sup>[31]</sup> 나가기 전, 해리는 알등도어에게 교수님은 이 거울을 보면 무엇이 보이냐고 물었고 알등도어는 자신이 이 거울을 보면 알등 한 칼자를 들고 있는 자신의 모습을 본다고 대답한다.<sup>[32]</sup>

여섯 후, 네빌이 알포이와 다리 루키 주위에 당한 케 그리핀으로 후계살까지 통통 튀어온다. 그런 네발을 위해서 루키 위해 헤르미온드가 크리스마스 선물로 준 것 중 마지막 개구리 초콜릿을 뜯은 해리는 금합영차 데저팅 알등도어를 함께 되는데, 그 뒷면의 내용에서 나폴라 콜라주를 발견하고, 그가 어딤사의 통통 만드는 유일한 연금술 사이어, 665살이라는 걸 알게 된다. 헤그리드가 713번 금고에서 꺼낸 것, 그린고츠의 도둑이 훔치려던 것, 복숭아가 지키고 있는 것은 불교상상의, 마법사의 통이었던 것이다.

2번째 허디지 경기, 우물부르피와 경기인데, 실은 스페이브였고, 알등도어 교수가 관전하러 나왔다. 그리핀도르는 스페이브의 편만 관성에도 불구하고 해리의 영활만으로 승리했다. 경기 후 해리는 어둠 속에서 크리스마스 선물로 받은 콜라주를 열어 보게 된다.<sup>[33]</sup>

시퀀스 길이는 셋은 헤그리드가 리커 콜드윈에서 만난 사형사지 카르타인으로 용의 말을 얻었다는 걸 알게 된다.<sup>[34]</sup> 용의 말은 곧 주된데, 헤그리드는 그 태어난 새끼용의 이름을 노버트로 짓는다.<sup>[35]</sup> 그러나 헤그리드의 집에서 용을 기르는 것은 불가능할 뿐더러,<sup>[36]</sup> 용이 노버트에게 손가락을 물리고 종족도어 입찰을 한 대다가<sup>[37]</sup> 알포이가 용을 피버렸기 때문에 노버트를 웨이너리에서 용을 연구하는 용의 왕의 위즐리에게 넘겨주기로 한다. 그러나 알리가 보낸 편지를 끼워 둔 용의 새끼용 알포이가 거지가버리는 바람에 그가 계획을 알아버린다. 해리 말탈은 알포이가 계획은 알도 투명 알도의 존재는 모른다. 정을 노려 아방에 투명 알도를 이용해 천문함으로 단견히 용을 배달하였다. 그러나 나무 산란 나이지 투명 알도를 두고 오는 바람에 울금 시간에 기숙사를 나간 것이 알려지고, 용에 대해 고자질하기 위해 메시나 방에 나온 알포이와 함께 해리, 헤르미온드, 그리고 알포이가 나갔다고 주의를 주러 나온 네빌까지 각각 기숙사 정수 50점의 도합 150점을 감점당하고 정계도 받게 된다.

정계는 헤그리드의 함께 금지된 숲에서 그의 말을 듣는 것이었다. 최후에 유니콘이 자주 죽는데 그 이유를 알아내는 것이 임무였다.<sup>[38]</sup> 처음에는 헤그리드, 해리, 헤르미온드가 한츠, 네빌, 알포이, 그리고 헤그리드의 개 같이 한츠로 움직였다. 그 외중에 금지된 숲에 헨타우로스도 만나 잠시 이야기도 하는데, 이때 네빌 쪽에서 이상을 의미하는 불꽃을 쏘아 올리고 헤그리드가 바로 달려가는데, 하지만 그건 알포이가 네빌을 놀리려는 장난을 치서 올린 네빌이 쏘아 올린거였고, 헤그리드는 네빌을 자신의 조에 넣고 해리를 알포이와 가계한다. 알포이와 이동을 하다 유니콘의 뒷자극을 보고, 그걸 따라 움직이던 해리는 어떤 무기를 쓴 알포이 유니콘의 피를 빨아먹는 것들을 본는데,<sup>[39]</sup> 동시에 이미의 후퇴에 엄청난 고통을 받는다. 다행히 헨타우로스 피탄체가 해리를 구해준 뒤 헤그리드에게 데려가 준다. 해리는 그린고츠의 도둑과 표를 붙이는 시간의 배후에 블드포트가 있다고 생각한다. 그리고 정실로 돌아가자, '만살을 대비하여'라는 해지와 함께 다시 투명 알도가 돌아와 있다.

그렇게 의욕에 빠진 상태에서 기말고사를 치르는 와중, 해리의 용다는 계속 아프고, 헤그리드에게 찾아가 용의 말을 쓴 사람에게 대해 이야기하는데<sup>[39]</sup> 알고 보니 그 사람은 처음부터 복숭아에게 지대한 관심을 갖고 있었고 결국 헤그리드가 슬기롭게 그에게 복숭이를 치내하는 방법<sup>[40]</sup>을 알려줬었다. 이를 알게 된 해리와 론, 헤르미온드는 그 사람이 스페이브나 블드포르라고 확실하게 어법사의 말을 출처를 할 것이라는 걸 알등도어 교수에게 알려주지만, 그는 어떤 영주의 초출을 받고 감시자를 할 한 상태였다. 대신 맥고나글 교수에게 알려주면 그녀는 그것은 안전하며, 또 다시 증거명령하면 또 기숙사 정수 50점을 감점하겠다고 한다.

해리, 론, 헤르미온드는 이방곳곳과 투명 알도를 쓰고 어법사의 통을 찾으려 가려고 하나, 네빌이 더 이상의 감정은 안 된다는 상충사를 기록한다. 헤르미온드가 급에 움직이던 주문으로 네발을 제압하고 서둘러 복숭이의 방으로 향한다. 헤그리드가 크리스마스 선물로 줬던 피리프 음악을 들려주자 복숭이가 잠들었고 지하실 문으로 떨어진 어지자 막사의 뒷안 서클에 흩어 단단히 묶이는데, 헤르미온드가 통을 열어 탈출한다.

Q:  
Huggingface에서 Context window가 4096인 Upstage-SOLAR 모델을 다운받았다.  
여기다가 10000토큰 문장을 truncation 없이 넣고 훈련이 가능할까?



# 용어 정리: context window란?

## “모델이 훈련할 때 사용한 시퀀스 길이”

셋은 콜라주에 대해 알아보려 여러 조사를 해 보나 나오지 않는다. 그 외중에 크리스마스가 나오고,<sup>[29]</sup> 해리는 크리스마스 선물로 헤그리드가 보낸 특이한 나무 리프도, 더글리 가족이 보낸 50원스 동전, 위즐리 부인이 보낸 달걀 피자와 스위트, 헤르미온드가 보낸 개구리 초콜릿을 들어 보고 마지막 선물을 찾는 데 발산이 있는 그 소도 안에서 어서와 의 것이라는 투명 알도를 선물받고, 이를 이용해 콜라주에 대해 찾아보려 도서관에 갔다 돌아오는 길에 헤매다가 한 방 안에 있는 거울을 보게 된다. 울 잠게도 그 거울을 보자 해리의 뒤에 부모님을 비롯한 가족들을 보고, 다음날 온과 함께 가지만, 온은 온과 회랑 테이블을 달고 기숙사 우승컵과 허디지 우승컵을 든 자신을 보게 된다. 그 거울에 빠져버린 해리는 또 찾아가는데, 이번엔 그 방에서 알도에게 교수들 만나게 되고, 그 거울이 현실의 자기 모습이 아니라 마음 속의 모습을 보여 주는 '소망의 거울'이라는 것을 알게 된다.<sup>[30]</sup> 그리고 다시는 거울을 찾지 않았다고 약속한다.<sup>[31]</sup> 나가기 전, 해리는 알도에게 교수님은 이 거울을 보면 무엇이 보이냐고 물었고 알도라는 자신이 이 거울을 보면 알뜰 한 할례를 물고 있는 자신의 모습을 본다고 대답한다.<sup>[32]</sup>

여섯 후, 네빌이 말포이와 다리, 루기 주위에 당한 케 그리핀으로 후계실까지 통통 튀어온다. 그런 네빌을 위로해 주기 위해 헤르미온드가 크리스마스 선물로 준 것 중 마지막 개구리 초콜릿을 들은 해리는 금형알차 대치할 알도에게 물게 되는데, 그 뒷면의 내용에서 나폴라 콜라주를 발견하고, 그가 어입사의 통을 만드는 유일한 연결을 사이며, 665살이라는 걸 알게 된다. 헤그리드가 717년 금고에서 꺼낸 것, 그린고츠의 도둑이 훔쳐버린 것, 복숭아가 지키고 있는 것은 불교왕상의, 마법사의 통이었던 것이다.

2번째 작디지 경기, 우물부르와의 경기인데, 실단은 스네이프로, 알도에게 교수가 관전하러 나왔다. 그리핀도르는 스네이프로 편마 관성에도 불구하고 해리의 영활만으로 승리했다. 경기 후 해리는 어두운

**Context Window**  
시퀀스 길이, 셋은 헤그리드가 리커 홀드업에서 만난 시퀀스화 커리저임으로 통의 통을 얻었다는 걸 알게 된다.<sup>[33]</sup> 통의 통은 주된하고, 헤그리드는 그 태어난 새끼통의 이름을 노버트로 짓는다.<sup>[34]</sup> 그러나 헤그리드의 집에서 통을 키르는 것은 불가능할 뿐이다.<sup>[35]</sup> 통이 노버트에게 손가락을 물리고 종족되어 입질을 한 대다가<sup>[36]</sup> 알도에게 통을 피바람이 때문에 노버트를 루나리아에서 통을 연구하는 통의 할 할의 위즐리에게 넘겨주게 된다. 그러나 할리가 보낸 편지를 끼워 둔 통의 책을 알도에게가자가 버리는 바람에 그가 계획을 알아버린다. 해리 말포이가 계획은 알도도 투명 알도의 존재는 모른다는 점을 노려 아방에 투명 알도를 이용해 천문함으로 단견히 통을 배달하였다. 그러나 나무 산남 나이지 투명 알도를 두고 오는 바람에 통금 시간이 기숙사를 나간 것이 알려지고, 통에 대해 고자질하기 위해 역시나 밤에 나온 알도이 함께 해리, 헤르미온드, 그리고 알도이 나갔다고 주의를 주러 나온 네빌까지 각각 기숙사 점수 50점씩 도합 150점을 감점당하고 징계도 받게 된다.

장기는 헤그리드의 함께 금지된 숲에서 그의 알을 통는 것이었다. 취급들이 유니콘이 자주 죽는데 그 이유를 알아내는 것이 임무였다.<sup>[37]</sup> 처음에는 헤그리드, 해리, 헤르미온드가 한츠, 네빌, 말포이, 그리고 헤그리드의 개 같이 한츠로 움직였다. 그 외중에 금지된 숲에 험터우르스도 만나 잠시 이야기도 하는데, 이때 네빌 쪽에서 이상을 의미하는 불꽃을 쏘아 올리고 헤그리드가 바로 달려가는데, 하지만 그건 알도이가 네빌을 놀리키는 장난을 쳐서 올린 네빌이 쏘아 올린거였고, 헤그리드는 네빌을 자신이 조에 넣고 해리를 말포이와 가계한다. 말포이와 이동을 하다 유니콘의 피타극을 보고, 그걸 따라 움직이던 해리는 어떤 무기를 쓴 할랑이 유니콘의 피를 빨아먹는 것쯤 보는데,<sup>[38]</sup> 동시에 이마의 후테에 엄청난 고통을 받는다. 다량히 험터우르스 피탄체가 해리를 구해준 뒤 헤그리드에게 대어가 준다. 해리는 그린고츠의 도둑과 표도를 붙이는 시간의 배후에 블드보트가 있다고 생각한다. 그리고 진실로 돌아가자, '만일을 대비하여'라는 해지와 함께 다시 투명 알도가 돌아와 있다.

그렇게 의욕에 빠진 상태에서 기말고사를 치르는 외중, 해리의 통리는 계속 아프고, 헤그리드에게 찾아가 통의 알을 쓴 사람에게 대해 이야기하는데<sup>[39]</sup> 알고 보니 그 사람은 처음부터 복숭아에게 지대한 관심을 갖고 있었고 결국 헤그리드가 술기운에 그에게 복숭아를 지나치는 방법<sup>[40]</sup>을 알려줬었다. 이를 알게 된 해리와 롬, 헤르미온드는 그 사람이 스네이프로나 블드보르라고 확실하게 어입사의 통을 훔치려 할 것이라는 걸 알면서도 교수에게 말하러안, 그놈 무의 통을 받고 감시처를 할 상해였다. 대신 맥고나글 교수에게 알리지만 그녀는 그것은 안전하며, 또 다시 장거행동하면 또 기숙사 점수 50점을 감점하겠다고 한다.

해리, 롬, 헤르미온드는 이랑곳없고 투명 알도를 쓰고 어입사의 통을 찾으려 가려고 하나, 네빌이 더 이상의 감정은 안 된다는 상충사를 기록하는데, 헤르미온드가 급세 움직이던 주문으로 네빌을 제압하고 서둘러 복숭아의 방으로 향한다. 헤그리드가 크리스마스 선물로 쥘턴 피리프 음악을 들려주자 복숭아가 감동했고 지하실 문으로 밀 어지자 막사의 뒷이만 석물에 흠이 단단히 묶이는데, 헤르미온드가 통을 훔쳐 탈출한다.

Q: Huggingface에서 Context window가 4096인 Upstage-SOLAR 모델을 다운받았다. 여기다가 10000토큰 문장을 truncation 없이 넣고 훈련이 가능할까?

A: 가능

즉, context window가 '길이 제한' 을 의미하지는 않음

# Sliding Window Attention

논문

## Longformer: The Long-Document Transformer

**Iz Beltagy\***   **Matthew E. Peters\***   **Arman Cohan\***  
Allen Institute for Artificial Intelligence, Seattle, WA, USA  
{beltagy, matthewp, armanc}@allenai.org

## Generating Long Sequences with Sparse Transformers

Rewon Child<sup>1</sup>   Scott Gray<sup>1</sup>   Alec Radford<sup>1</sup>   Ilya Sutskever<sup>1</sup>

모델



Mistral 초기 모델에만 들어가 있고,  
요즘 모델에는 사용되지 않음!

# Sliding Window Attention

Attend할 수 있는 토큰 개수를  
[window\_size]개로 제한!  
(예시 window\_size=4)



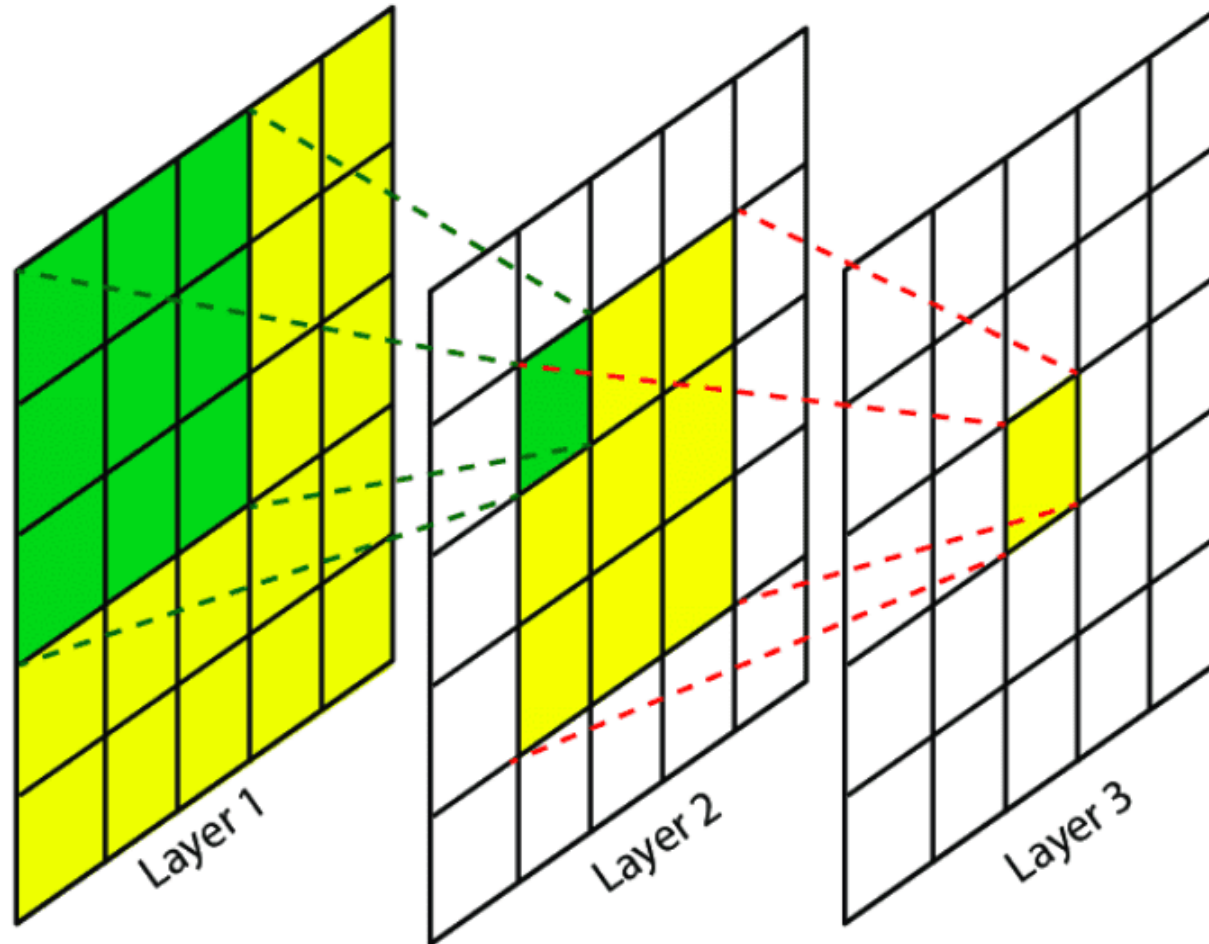
Layers



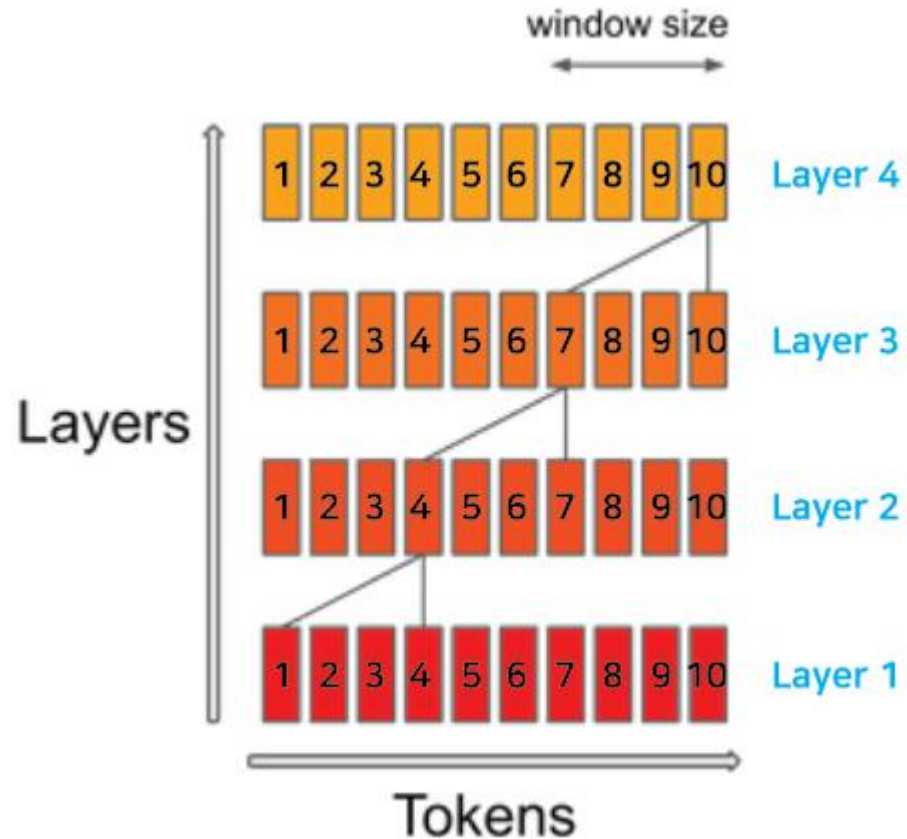
결국 Layer 4의 10번 토큰은  
Layer 1의 1~10번 토큰에  
모두 영향을 받게 됨

# Sliding Window Attention

CNN의 Receptive Field  
개념과 비슷하다.



# Sliding Window Attention 문제



- Global Attention (모든 토큰을 전부 보는 방식)에 비해 성능은...?
- 메모리의 효율성은...?

# 논문 1: Longformer: The long-document transformer

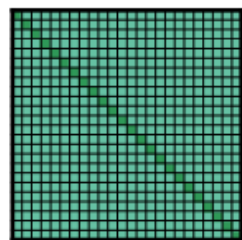
## **Longformer: The Long-Document Transformer**

**Iz Beltagy\***    **Matthew E. Peters\***    **Arman Cohan\***  
Allen Institute for Artificial Intelligence, Seattle, WA, USA  
{beltagy, matthewp, armanc}@allenai.org

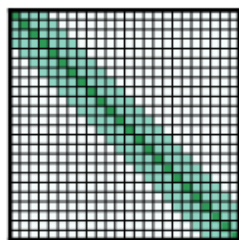
Sliding window attention의 아이디어를 LM에서 구체화한 논문  
사실 아이디어는 두 번째 논문이 다 하긴 했지만...

그래도 Language model을 타겟으로 구체화해서 그런지, 인용 수가 많음  
(3700+ 회)

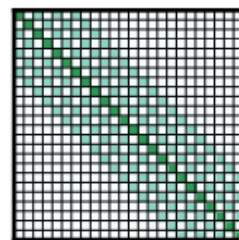
# 논문 1: Longformer: The long-document transformer



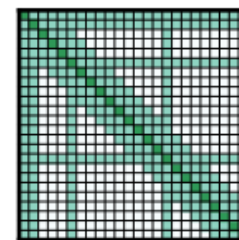
(a) Full  $n^2$  attention



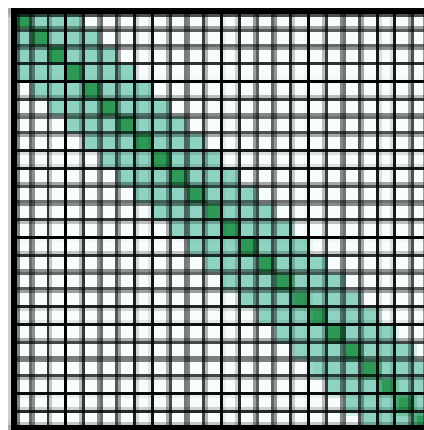
(b) Sliding window attention



(c) Dilated sliding window



(d) Global+sliding window



(b) Sliding window attention

# 논문 1: Longformer: The long-document transformer

Model	Dev BPC
Decreasing $w$ (from 512 to 32)	1.24
Fixed $w$ (= 230)	1.23
Increasing $w$ (from 32 to 512)	<b>1.21</b>
No Dilation	1.21
Dilation on 2 heads	<b>1.20</b>

Table 4: Top: changing window size across layers. Bottom: with/without dilation (@ 150K steps on phase1)



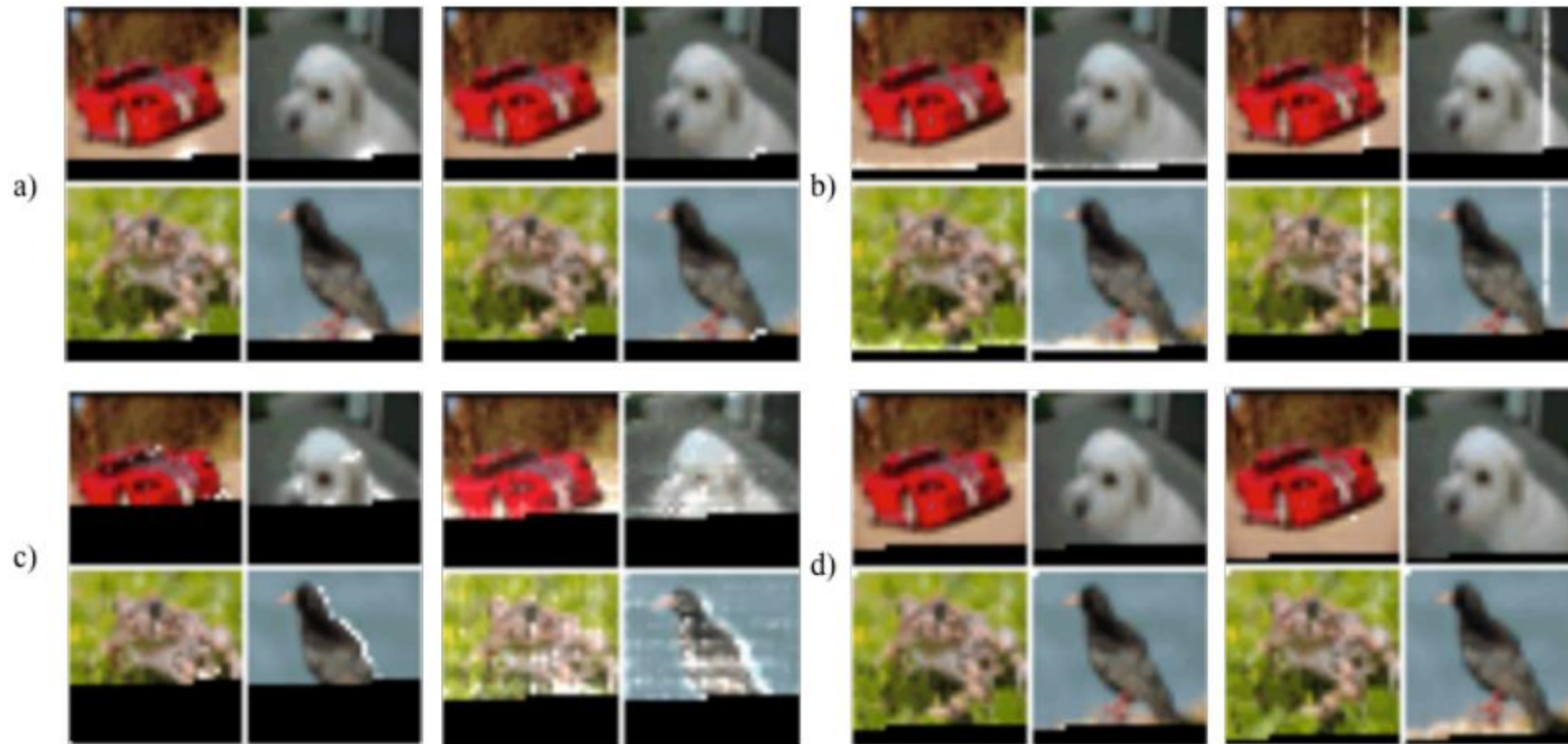
## 논문 2: Generating Long Sequences with Sparse Transformers

<b>Generating Long Sequences with Sparse Transformers</b>
Rewon Child <sup>1</sup> Scott Gray <sup>1</sup> Alec Radford <sup>1</sup> Ilya Sutskever <sup>1</sup>

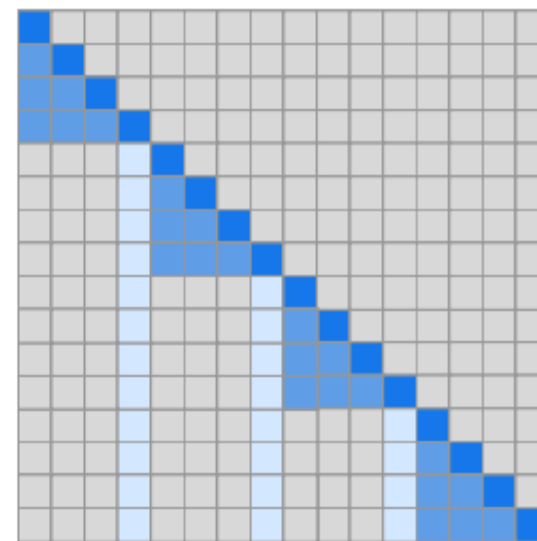
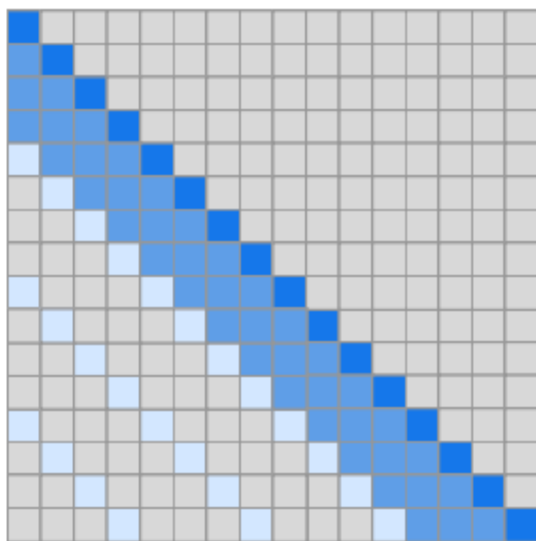
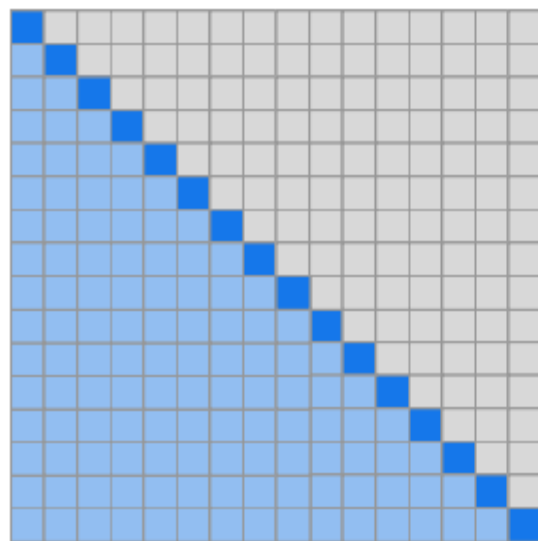
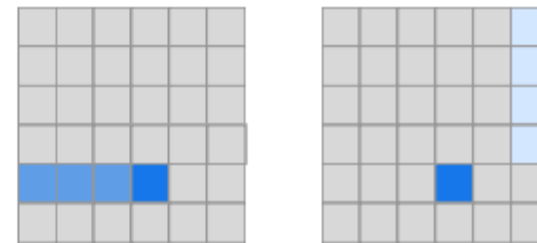
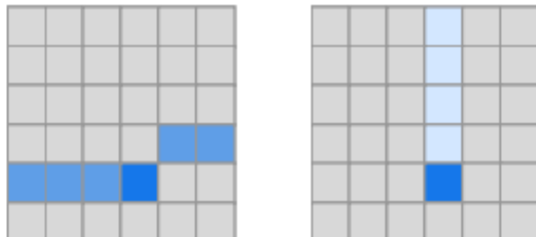
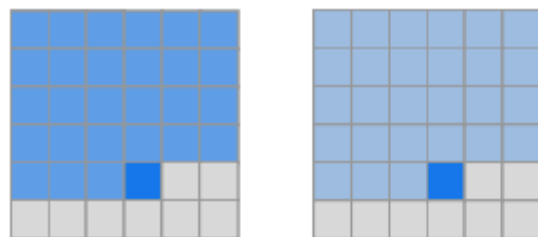
Sliding window attention의 아이디어를 일반적으로 실험한 논문  
논문이 매우 짧고 간결하다. 장황한 말 대신 수식과 예제로 설명한다.  
이해하는데 시간이 좀 걸린다. 성의 없게 그린 그림을 이해해야 한다.

얘는 OpenAI 논문이고, 비전 모델까지 모두 아우르는 연구라서 그런지 인용 수가 많다.  
(1680+ 회)

## 논문 2: Generating Long Sequences with Sparse Transformers



## 논문 2: Generating Long Sequences with Sparse Transformers



(a) Transformer

(b) Sparse Transformer (strided)

(c) Sparse Transformer (fixed)

그러나... SWA 없이 힘으로 밀어붙이는 추세

- Mistral NeMo: 128K, No sliding window
- Command R+: 128K, No sliding window
- Llama 3.1: 128K, No sliding window

대신 H100 80G가 10000장 이상 들었다는 것!

# Long sequence 처리와 관련있는 기술들

- RoPE (Rotary Position Embedding)
- GQA (Grouped Query Attention)
  - KV-cache가 왜 필요하고, 어떨 때 쓰이는지 이해하면 좋음
- Ring Attention
  - Gemini-1.5가 어떻게 1.5M context window를 달성할 수 있었는가?
  - SWA를 제외하면 가장 유력한 방법... 정확하게는 알 수 없음.