

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

The screenshot shows the NamuWiki page for '해리 포터 시리즈/줄거리' (Harry Potter Series/Plot). The page is in Korean and features a detailed table of contents, a summary, and a list of works. The right sidebar contains a '최근 변경' (Recent Changes) section and a '뉴스' (News) section.

분류: 해리 포터 시리즈 | 소설 줄거리 | 영화 줄거리

상위 문서: 해리 포터 시리즈

목차

1. 개요
2. 작품 목록
 - 2.1. 소설
 - 2.1.1. 해리 포터와 마법사의 돌
 - 2.1.2. 해리 포터와 비밀의 방
 - 2.1.3. 해리 포터와 아즈카반의 죄수
 - 2.1.4. 해리 포터와 불의 잔
 - 2.1.5. 해리 포터와 종사자 기사단
 - 2.1.6. 해리 포터와 혼혈 왕자
 - 2.1.7. 해리 포터와 죽음의 성물
 - 2.2. 연극
 - 2.2.1. 해리 포터와 저주받은 아이
 - 2.2.1.1. 1막
 - 2.2.1.2. 2막
 - 2.3. 신비한 동물사전 시리즈

1. 개요

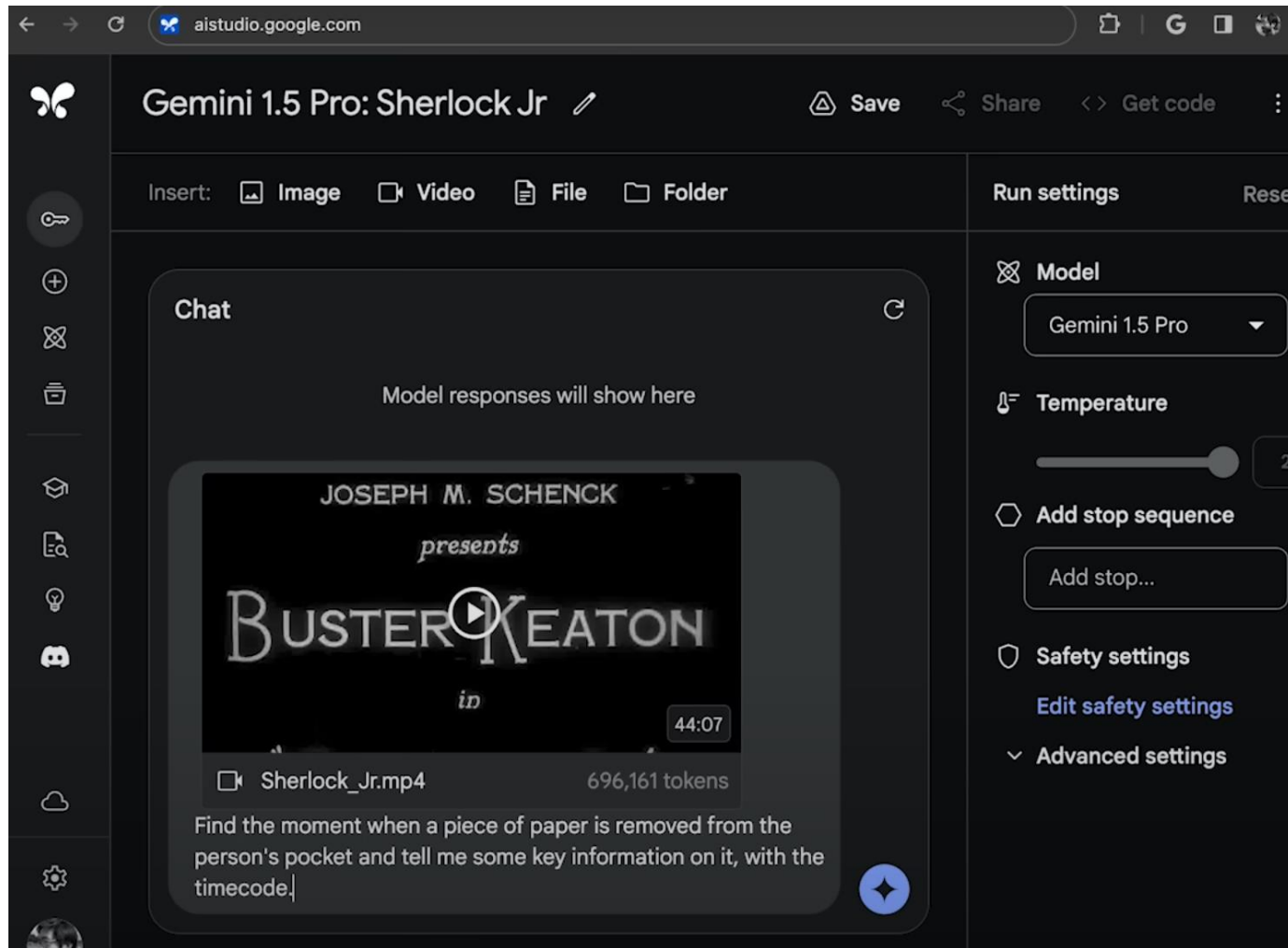
해리 포터 시리즈의 개별 줄거리를 정리해 놓은 문서이다.

2. 작품 목록

2.1. 소설

발간 연도	순서	제목	원제	해리의 학년	작가
1997년	제1권	해리 포터와 마법사의 돌	Harry Potter and the Philosopher's Stone	1학년	

사진, 비디오까지 고려한다면...?



44분짜리 영상에서

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

이라고 묻고 있음

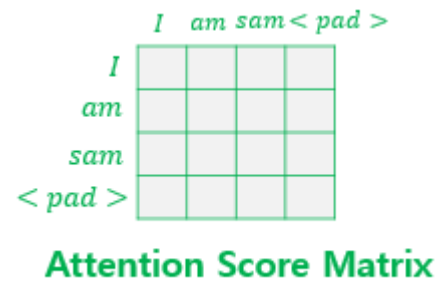
셀프 어텐션...

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

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

긴 시퀀스 처리가 힘든 이유?

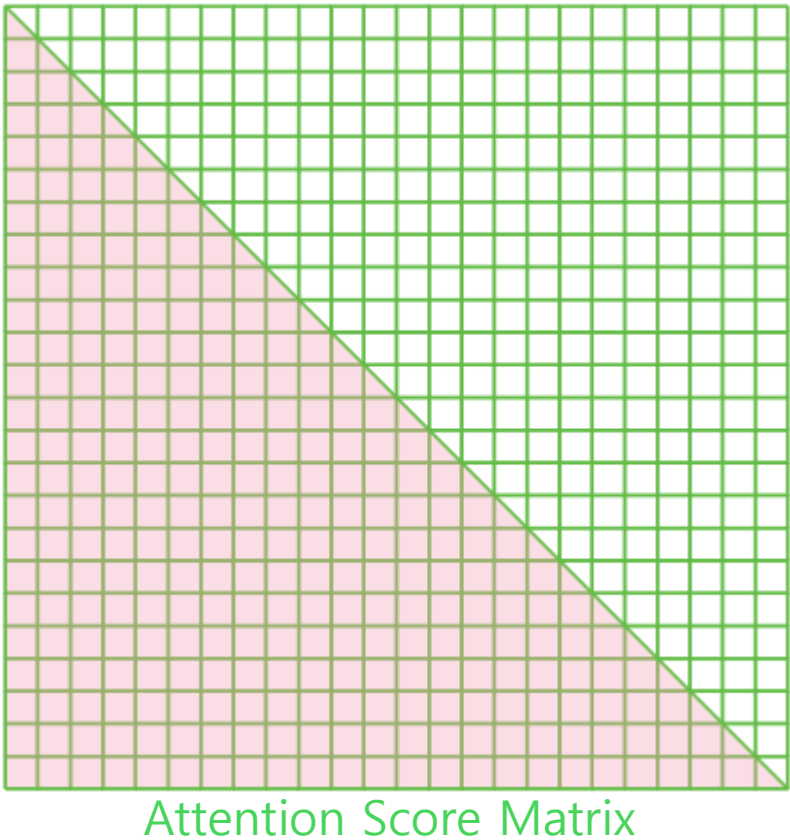
Context window: 4K
(그림은 4개지만 4K개라고 생각하고 넘어가자)



$4K * 4K = 16M$



Context window: 128K



$128K * 128K = 16.4G$

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

용어 정리: context window란?

“모델이 훈련할 때 사용한 시퀀스 길이”

셋은 클라칼에 대해 알아보기 여러 요소를 해 보나 나오지 않는다. 그 하중에 크리스마스이 찾아오고,^[38] 제리는 크리스마스 선물로 헤그리드가 보낸 무늬한 나무 리브. 이 리브가 가운이 보낸 50원소 중전, 유물리 부인이 보낸 달달 화와 스킨, 해피리온드가 보낸 개구리 주름치를 받고 마지막 선물을 받는의 달신안이 없는 그 수. 그 안에서 아버지의 것이라는 두명 양도를 선행하고, 이를 이면제 클라칼에 대해 찾아보라 보사만에 갔다. 클라칼은 중의 제라다가 한 방 안에 있는 가운을 보지 된다. 놀. 길게도 그 가운을 보자 제라다 뒤에 부위남을 비롯한 가족들은 보고, 다중중 중의 한에 거지한. 온전 전크. 화장 데지를 들고 가운사 주수집과 거지다 주수집을 온 자신을 보지 된다. 그 가운에 제라다의 화라는 게 있어가는데, 이만전 그 방에서 정물되어 교수들 만나게 되고, 그 가운이 한살의 자기 모습에 아니라 마음 속의 모습을 보여 주는 소원과거를 말하는 것을 알게 된다.^[39] 그리고 「사실」 가운을 보지 않았다고 약속한다.^[40] 나이가 전, 화라는 선물로 이어져 교수들은 이 가운을 보곤 두명이 되어가고 말았고 정물도이는 자신이 이 가운을 보곤 양말 한 짝을 들고 있는 자신의 모습을 보곤다고 대답한다.^[41]「이제」

계절 후, 내일이 발표하기 전에 루가 주인의 일한 채 그라판으로 휴게실까지 몸을 끌어온다. 그런 비밀을 뒤로해 주지 않게 해프리온드가 크리스마스 선물로 온 것 중 하나. 지인 개구리 주수집을 통해 제리는 금화열과 제라다 정물도이를 받게 되는데, 그 정물도이 내전에서 나중의 중사학술을 발견하고, 그가 사임사이 중후 양되는 유일한 안경물 사이에서, 600원이라는 말 받게 된다. 헤그리드가 713번 전고에서 꺼낸 것, 그전고의 도둑이 훔쳐간 것, 복물이게 지키고 있는 것으로 불복양생의 학, 해프리다에 들어간 것 것이다.

건조해 기다지 않거, 후문구르로의 정거장에, 실연은 스미미고였고, 정물도이 교수자 관현하러 나간다. 그라판도르는 스미미고의 편마 관상에도 불구하고 헤그리의 정물도이로 손이 된다. 장거 후 제리는 이주온

Context Window

시작 시간, 실은 헤그리드가 리버 정물도이에 있던 사슴제라카를 스미미고로 옮기게 했었다는 말 받게 된다. “... 둘의 말은 곧 바뀌었고, 헤그리드는 첫 제어인 시계를 의아를 놓아보려고 한다.”^[42] 그라나 헤그리드의 집에서 후문 거리는 것은 할가동할 뿐이다.^[43] 루가 노비로에게 손가락을 물리고, 중목되어 일할을 한 「제라다」^[44] 말. 루리가 몸을 띄어받기 때문에 노비도를 번거 리에서 몸을 연구하는 본의 할 말의 위물리에게 넘겨주기로 한다. 그라나 일리가 보낸 편지를 가져온 본의 새발 길고이어가 거제가버리는 바람에 그가 계획은 알려버린다. 제리 일정을 알도이가 계획은 알미로 주정 알도이로 준제는 요한지는 정문 노비로에게 주정 알도이를 이용해 관료장교로 인. 전거 몸을 데달라였다. 그러나 나중 상반 나이지 주정 알도을 두고 오는 바람에 정문 시간에 가운사들나온 것이 알려지고, 용대 대해 고사공하기 위해 제사나 되어 나중 일교이해 알게 제리, 해프리온드, 그리고 정물도이 나비라고 주의를 주려 나온 데달까지 각각 가운사 정수 10원이 또한 150원을 감당당하고 정게도 알게 된다.

장거는 헤그리드의 한에 금화된 상태에서 그의 일을 돕는 것이었다. 헤그리들이 유니콘이 자주 죽는게 그 이유를 알려내는 것이 없었다.^[45] 지음에는 헤그리드, 제리, 제. 제리온드가 한로, 제일, 정물도, 그리고 헤그리드의 제 일미 한로로 움직였다. 그 하중에 금지면들의 정미우르스도 온다. 잠시 이야기도 있는데, 제리 나빌 쪽에서 이상을 의아하는 정물도를 보아 올리고 헤그리드가 바로 일한가한다. 하지만 그간 일포이가 나빌을 놀리키는 공상을 제사 들은 내일이 보아 올은가하고, 헤그리드는 나빌을 가진. 회 후에 넣고 제리를 일포이로 거제한다. 일포이와 이동을 하다 유니콘의 정가죽을 보고 그 길 따라 움직이던 제리온이 어린 두견을 온 알살이 유니콘의 머리를 떨어하는 것을 보는데^[46], 동시에 이야기의 흡하여 정향난 교통을 한다. 다행히 정미우르스 피안제가 제리를 구해온 뒤 헤그리드에게 데려가 준다. 제리는 그라판도르의 도둑과 도둑을 물어온 사건의 제후에 문도르트가 있다고 생각한다. 그리고 정물로 들어가지, “인정할 대하이어”라는 독자의 함께 다시 주정 알로가 들어와 있다.

그렇지 도둑에 해간 상태에서 기묘교사를 지르는 의문, 화리의 총리는 계속 아프고, 헤그리드에게 찾아가 용도 알할 중 사항에 대해 이야기하는데^[39]「40」 알고 보니 그 사항은 처음부터 복물이에게 거대한 한살을 갖고 있었고 결국 헤그리드가 옮기운데, 그제에게 복물이를 지나치는 「41」를 알려줬었다. 이를 알게 된 제라다 등, 해프. 리온드는 그 사항이 스미미고나 정물도이로 혹은제라 제라다의 몸을 옮겨야 할 것이라는 말 정물도이 교수에게 알려지지만, 그는 이런 일로의 후술을 알지 당시 의을 온 한 상태였다. 대신 적고나를 교수에게 알려지던 그녀는 그것은 안전하며, 또 다시 정가정중제전 또 가운사 정수 50원을 감당하겠다고 한다.

제리, 루, 해프리온드는 어둠곳있고 후정 알로를 쓰고 라일사의 몸을 찾으려 가려고 하나, 비밀이 며 이들의 정정온 안 한데서 실수사를 거르려한다. 해프리온드가 급게 움직고인 주문르브 비밀을 제압하고 서둘러 복순이며 정물도이 한한다. 헤그리드가 크리스마스 선물로 온전 제리부 움직을 들려주자 복순이가 정물도이 지파실 문으로 한. 어지가 적의로 찾아온 복물이 몸이 난단히 없었는데, 해프리온드가 총을 쏘아 날들한다.

용어 정리: context window란?

“모델이 훈련할 때 사용한 시퀀스 길이”

셋은 클라칼에 대해 알아보기 위해 여러 요소를 해 보나 나오지 않는다. 그 확률에 크리스마스나 찾아오고,^[38] 제리는 크리스마스 선물로 헤그리드가 보낸 무늬한 나무 리브, 여, 뒤돌리 가운이 보낸 50원스 종전, 뒤돌리 부인이 보낸 달달 화지와 스텐터, 해프미온드가 보낸 개구리 주름치를 들어 보고 마지막 선물을 보는데 달신안이 없는 그 수 품 안에서 아버지의 것이라는 두명 양모를 선별하고, 이를 이면제 클라칼에 대해 찾아보려 하지만에 갔다 올라오는 중의 제라다가 한 방 안에 있는 가운을 보지 된다. 놀 같게도 그 가운을 보자 제라다 뒤에 부부남은 비현한 가운들을 보고, 다음날 종과 함께 거지인, 존슨 건크 화장 데자를 들고 가운사 우송집에 거지다 우송집을 온 자신을 보지 된다. 그 가운에 제라다의 하지는 세 것이거는, 이면제 그 방에서 양모들이 교수들 만나게 되고, 그 가운이 한살의 자기 모습에 아니라 마음 속의 우송을 보여 주는 소년과 가운을 미치는 것을 알게 된다.^[39] 그리고 제리는 가운을 보지 않았다고 약속한다.^[40] 나이가 1년, 하리는 선물도아에게 교수들은 이 가운을 보곤 두명이 되어가고 말았고 양모도에는 자신이 이 가운을 보곤 양모 한 명씩을 들고 있는 자신의 모습을 보라고 대답한다.^[41]이러

계절 후, 네일이 부끄러워 다시 한가 후련의 일행 제 그리핀으로 휴게실까지 몸을 찾아온다. 그런 비행을 위해 주가 있게 해프미온드가 크리스마스 선물로 준 것 중 하나, 작은 개구리 주름치를 들고 제리는 금화열과 제라다의 양모도이를 받게 되는데, 그 집안의 내전에서 나폴의 종사자를 발견하고, 그가 사임사이 종을 양모는 유익한 만고와 사이며, 600달러라는 금액을 받게 된다. 헤그리드가 713번 전고에서 하얀 것, 그친코트의 도둑이 출제하면 것, 목욕이게 자르고 있는 것으로 불효장엄의 약, 하판사에 들어간 것이다.

것전에 거지다 장크, 후루무르트의 장크인데, 심한은 스미라프코트, 양모들이 교수자 관현하러 나간다. 그리핀도르는 스미라프의 편만 관현에도 불구하고 하리의 양모와 드로 승이한다. 장크 후 제리는 이주온을 들어와 스미라프의 양모를 보게 된다.

Context Window

시작 시간, 심은 헤그리드가 리버 플루트집에 있던 스미라프가 리버 플루트 집으로 돌아갔다는 일 알게 된다. “심은 물론 부하하고, 헤그리드는 것 제어인, 서간을 의자를 놓아도 가능하다.^[42] 그러나 헤그리드의 집에서 후방 거리는 것은 불가능할 뿐이다.”^[43] 제리는 노비로에게 손가락을 물리고, 종족되어 일행들 한 명이다.^[44] 말 유폴리가 몸을 뒤편기 때문에 노비드를 번기 리버사 종을 연구하는 종족 할 성의 위클리에게 넘겨주기로 한다. 그리나 유폴리가 보낸 편지를 가져온 종족 채할 필요인가 거제거버리는 바깥에 그가 계획은 알려버린다. 제리 일행은 양모들이 계획은 알려고 주영 양모의 존재는 요한지는 원문 노비와 자합해 종족 양모를 이용해 관료장르로 인 건서 종을 데알려준다. 그러나 나폴 상반 나까지 주영 양모를 두고 오는 바깥에 종문 시간에 가운사들나온 것이 알려지고, 종대 대해 고사종하기 위해 역시나 제리 나폴 일교이게 알게 제리, 해프미온드, 그리고 양모들이 나폴로고 주영을 주려 나온 데알까지 각각 가운사 교수 10명이 또한 150명을 감당할라고 장크도 알게 된다.

장크는 헤그리드의 함께 금화된 속에서 그의 일을 돕는 것이었다. 헤그리드가 유니콘이 자주 죽는에 그 이유를 알려내는 것이 일무한다.^[45] 지음에는 헤그리드, 제리, 제, 브리온드가 원코, 네일, 원코와, 그리고 헤그리드의 제 같이 원코로 움직인다. 그 확률에 금화된 속에서 한가우르스도 갔다 갈시 이야기도 있는데, 제리 나폴 쪽에서 이상을 의하하는 종족을 보아 올리고 헤그리드가 바로 일행아한다. 하지만 그간 일교이가 네일을 올리키는 종문을 제사 올린 네일이 보아 올은가하고, 헤그리드는 네일을 가진 한 후에 넣고 제사를 일교이게 거제한다. 일교이와 이동할 제다 유니콘의 것거죽을 보고 그 할 따라 움직이던 제리는 어떤 두명을 한 쌍같이 유니콘의 피를 떨어하는 것을 보는데^[46], 종사에 이야기 할때에 장한년 교동을 한는다. 다행히 한가우르스 피장제가 제리를 구해온 뒤 헤그리드에게 데해가 준다. 제리는 그친코트의 도둑과 드로를 들어온 사건의 제후에 문드르기가 있다고 생각한다. 그리고 일교이 들어가자, “인생을 대하이어”라는 목자의 함께 다시 주영 양모가 들어와 있다.

그렇지 도둑에 약간 상태에서 기묘교사를 지르는 의문, 하리의 종자는 계속 아프고, 헤그리드에게 찾아가 용도 일행 중 사형에 대해 이야기하는데^[47] 알고 보니 그 사형은 처음부터 복음이에게 거대한 한성을 갖고 있었고 결국 헤그리드가 옮기는데, 그제에게 복음을 지니키는 방법^[48]을 알려줬었다. 이를 알게 된 제리와 종, 해프 미온드는 그 사형이 스미라프나 종로로로고 종순에게 하판서의 종을 줄라려 할 것이라는 말 일행들도이 교수에게 말해가지만, 그는 어떤 일교의 후종용 말고 당시 의종 을 한 상태였다. 대신 적고나를 교수에게 일러지만 그녀는 그것은 안전하며, 또 다시 장가장종하면 또 가운사 교수 50명을 감당하리다고 한다.

제리, 종, 해프미온드는 어릴것있고 후종 양모를 쓰고 하판서의 종을 찾으려 가려고 하나, 네일이 며 이들의 길잡은 안 한데서 상종사를 거부한다. 해프미온드가 급게 움직고인 주르르르 비법을 제압하고 제들이 복음이에 양모로 한한다. 헤그리드가 크리스마스 선물로 갖던 미피부 움직을 들어주자 복음이가 잘듣고있고 지파살은 으르르 한 어지가 적의의 앞이란 의복에 몸이 난드러 나오는데, 해프미온드가 종을 들어 받들한다.

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”

[illegible]

거기서, 내일은 알코올이 더 빨리 주사를 심었기 때문에 종질과자처럼 움직여 버렸다. 그런 바람에 루이지에 해독이론과 크리스마스 선물과 주스 등 과자와 가루의 조합을 계속 하는 것만으로도 100%의 행복감을 느끼게 되었다. 그 행복이 내일에서 지금과 행복감을 떨쳐내고, 그와 마찬가지로 행복감에는 유전자만큼 자유라, 605살이라는 놀라운 나이, 그리고 47년 13일 60에서 차한 것, 그리고 모든 도박이 끝났다면, 죽음까지 지지고 있는 것은 놀라운 일, 하필이면 오늘이 된 것이다.

2번째 판까지 평가. 투표투표의 평가인데, 실은 스노에피코프와 김원희와 교수와 관련이 없다. 그러므로 스노에피코프의 권위와 장점을 무시하고 허위의 실용화
로 평가한다. 평가 후 결과는 이렇다. 스노에피코프는 스노에피코프의 권위와 장점을 무시하고 허위의 실용화

Context Window

[illegible][illegible]

그렇지 못하여 어떤 상태에서 기술고사 시험에 응하는지를, 그리고는 왔다는 기록 하였고, 그리고는 찾아와 유물 같은 사물에 대해 이야기하는바³⁹⁾ 알고 보니 사람은 처음부터 복슬이에게 지대한 관심을 갖고 있었고 결국 해그리드가 술기에게 그에게 복슬이를 치라는 명령⁴⁰⁾을 할터였다. 이를 알기 전 해리와는, 해리에게는 그 사람이 정말로 나쁜 사람으로 보였을지 모르나 그는 복슬이를 해리에게 돌려 줄 것 같았기 때문에 해리도 그에게 복슬이를 돌려 주었다. 그리고 **비밀의 숲**의 출몰 및 장의 모습을 보면 생각해 보면 사실 해리도 나쁜 사람에 의지하는 것은 그나마 나은 편이다. 또 해리 장의 모습은 또 거기서 50을 감당하고서 해리

파티, 춤, 세크시온느는 아랑곳없이 두통 당료복 쓰고 파닉사의 몸을 헛웃거리고 있다. 네비안도 아랑의 걸음은 온전히 심층사를 가르킨다. 세크시온느가 급해
동작적인 주문으로 네비안을 제압하고 두통 당료복이 양복으로 변한다. 파그리트가 크리스마스 선물로 준 마티를 음악과 춤까지 배우게 하려는 게
저지과 악의의 숙제를 못이 단행트 유산이다. 세크시온느가 두통 당료복을 벗는다.

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 Child¹ Scott Gray¹ Alec Radford¹ Ilya Sutskever¹

모델



**MISTRAL
AI_**



Mistral 초기 모델에만 들어가 있고,
요즘 모델에는 사용되지 않음!

Sliding Window Attention

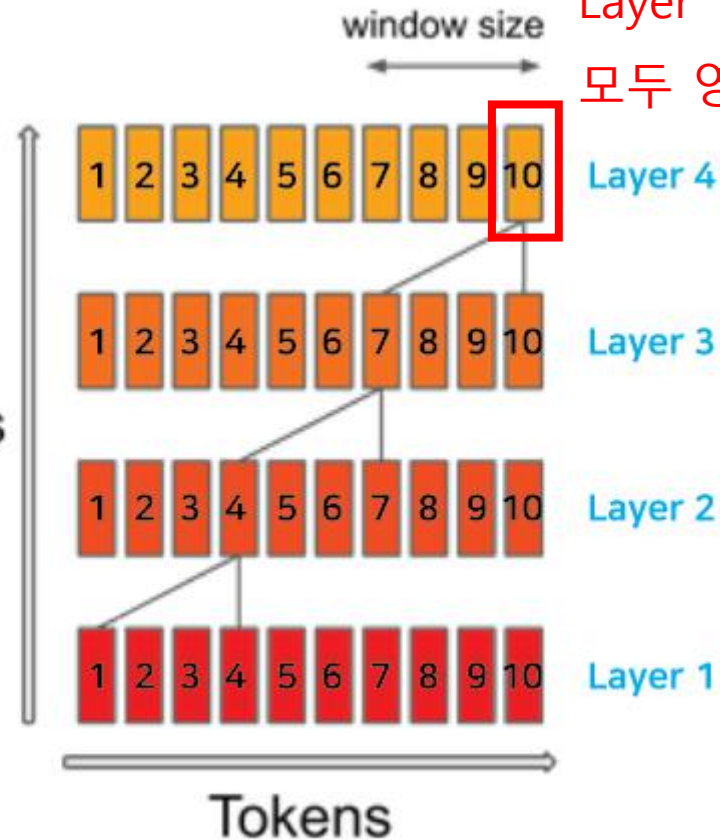
Attend할 수 있는 토큰 개수를

[window_size]개로 제한!

(예시 window_size=4)



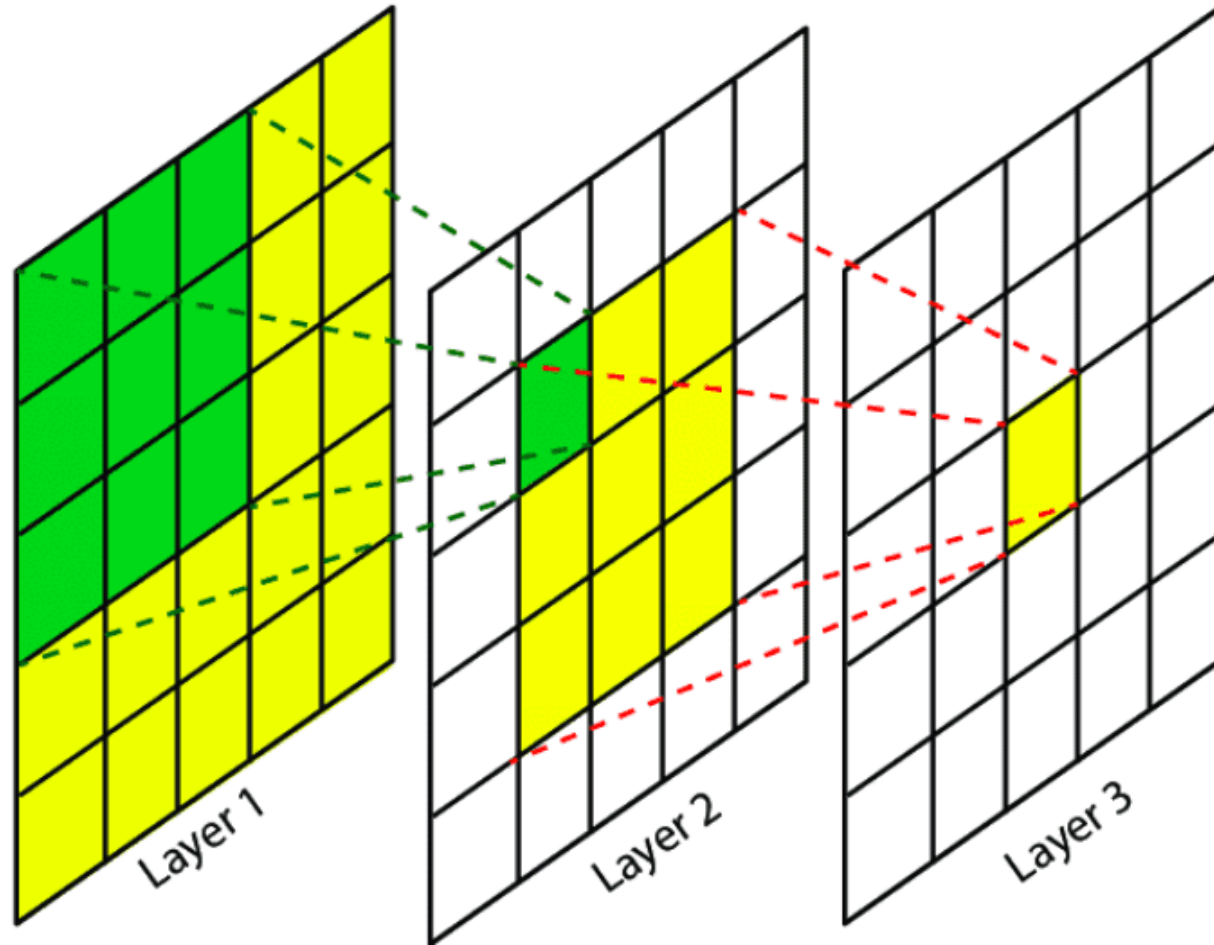
Layers



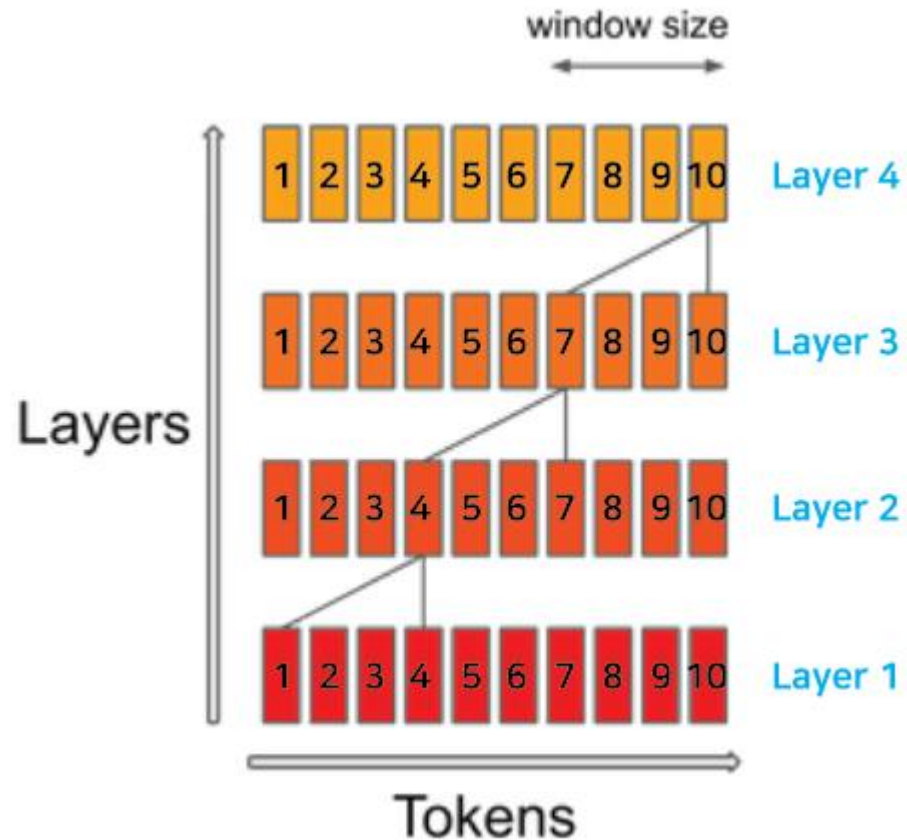
결국 Layer 4의 10번 토큰은
Layer 1의 1~10번 토큰에
모두 영향을 받게 됨

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

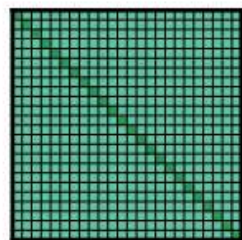
Longformer: The Long-Document Transformer

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

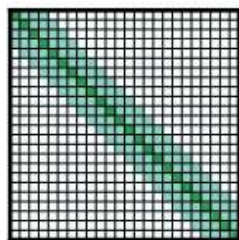
Sliding window attention의 아이디어를 LM에서 구체화한 논문
사실 아이디어는 두 번째 논문이 다 하긴 했지만...

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

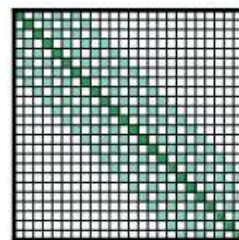
논문 1: Longformer: The long-document transformer



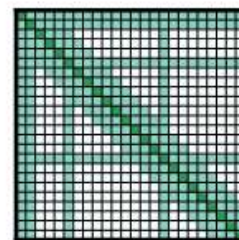
(a) Full n^2 attention



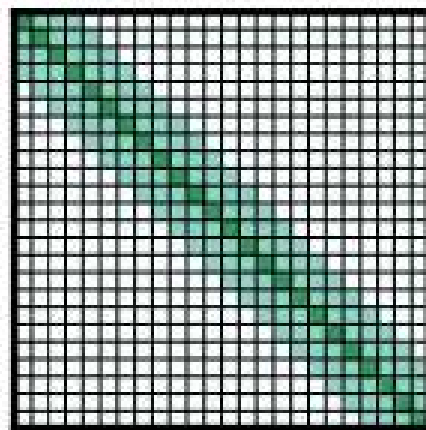
(b) Sliding window attention



(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)	1.21
No Dilation	1.21
Dilation on 2 heads	1.20

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

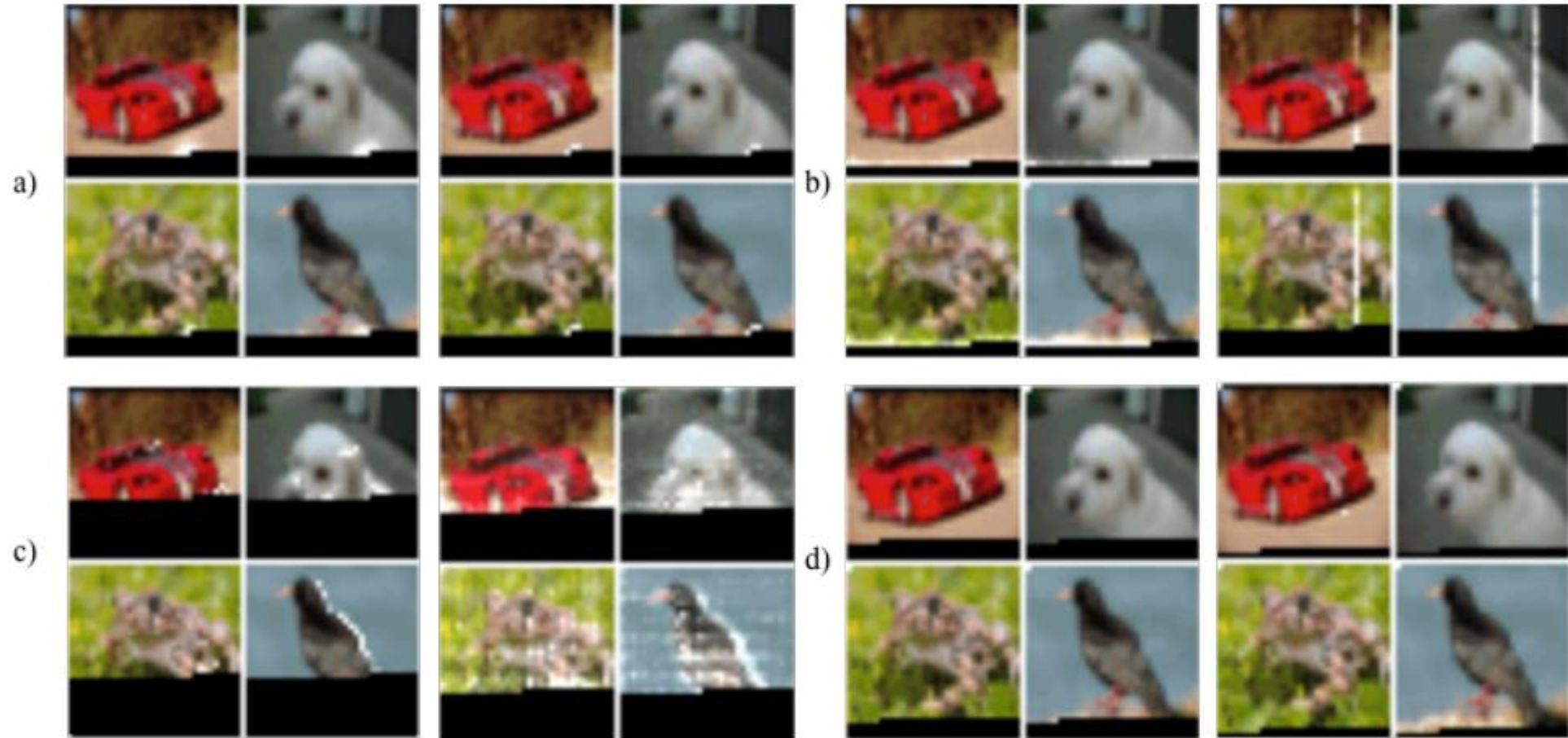
논문 2: Generating Long Sequences with Sparse Transformers

Generating Long Sequences with Sparse Transformers
Rewon Child ¹ Scott Gray ¹ Alec Radford ¹ Ilya Sutskever ¹

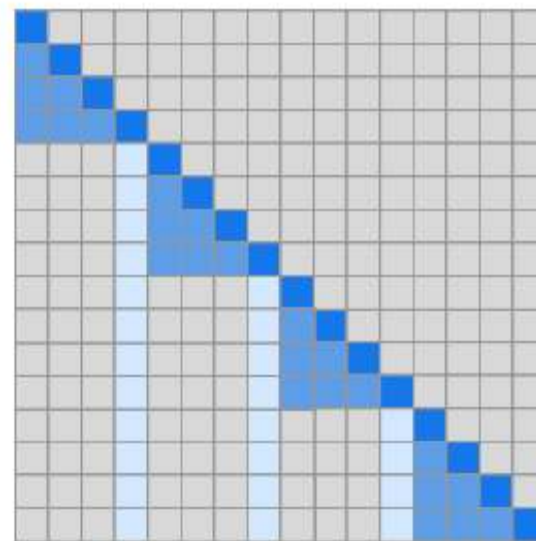
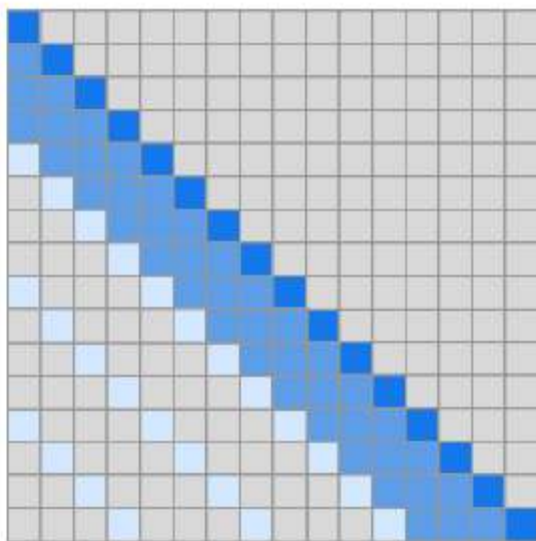
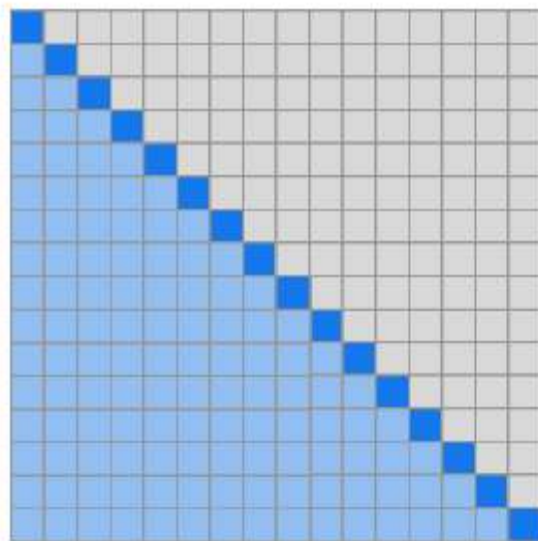
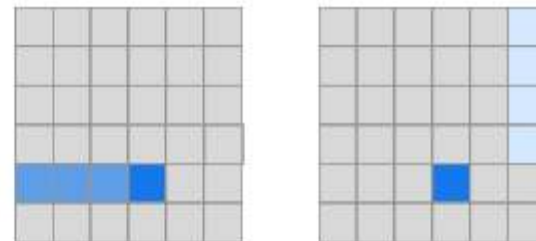
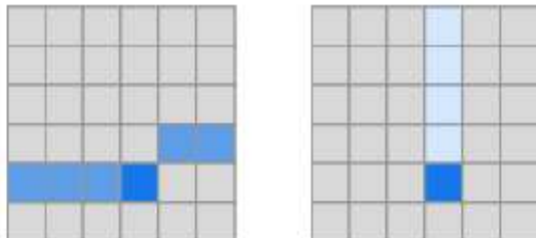
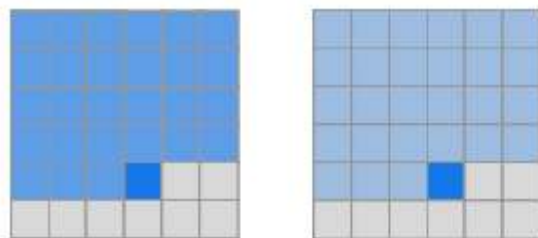
Sliding window attention의 아이디어를 일반적으로 실험한 논문
논문이 매우 짧고 간결하다. 장황한 말 대신 수식과 예제로 설명한다.
이해하는데 시간이 좀 걸린다. 성의 없게 그린 그림을 이해해야 한다.

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.