Massive Activation in transformer architecture



Pusan National University
Dept. of Electronic Engineering
Computer Vision & Signal Processing Lab.



Presented By
202583323 백창우

Outline

- 선정 이유
- 주제 소개
- Background
- Paper 1: Vision Transformers Need Registers (ICLR 2024 Oral)
- Paper 2: Efficient Streaming Language Models with Attention Sinks (ICLR 2024)
- Conclusion

선정 이유

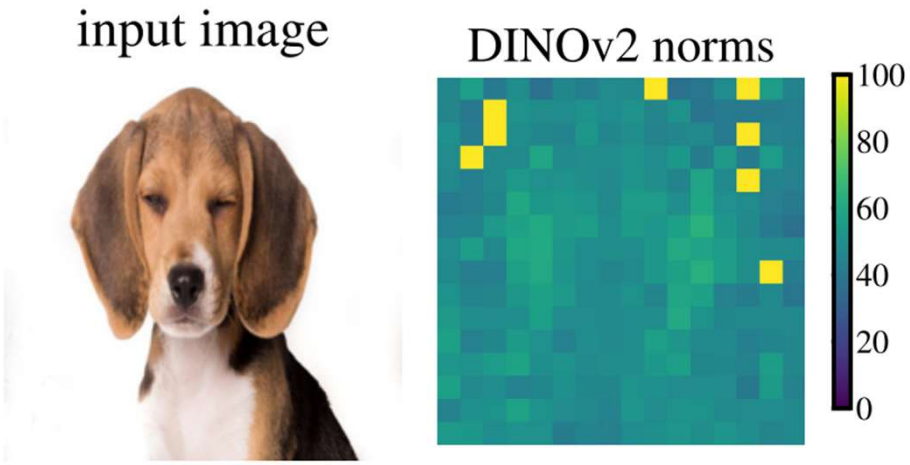
- 본인 연구와 긴밀히 연관 됨
  - Token Pruning

- 특정 메소드가 아닌 대형 Transformer 아키텍처 전반에서 관찰된 고유한 현상
  - 자연어, 이미지 등 모달리티를 구분하지 않음
  - Transformer를 활용하는 다른 연구자들에게도 큰 도움이 될 것

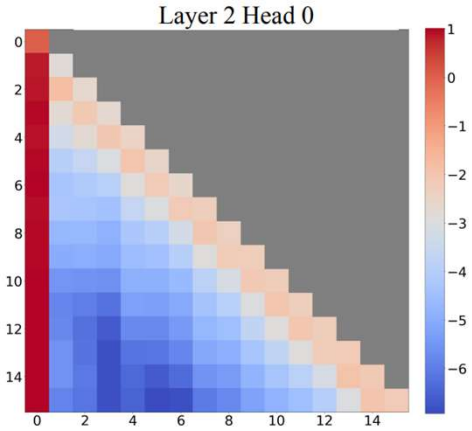
- 영향력 있는 논문 위주로 선정
  - Vision Transformers Need Registers (ICLR 2024 Oral)
    - 인용 수 531
  - Efficient Streaming Language Models with Attention Sinks (ICLR 2024)
    - 인용 수 826

주제 소개

- Massive Activation in transforemer
  - ViT나 LLM과 같은 대규모 Transformer 계열 모델에서, 의미 없는 토큰들이 비정상적으로 큰 norm 값을 갖는 현상이 보고됨
    - 이러한 토큰은 다른 토큰들로부터 과도한 attention value를 받는 경향이 존재



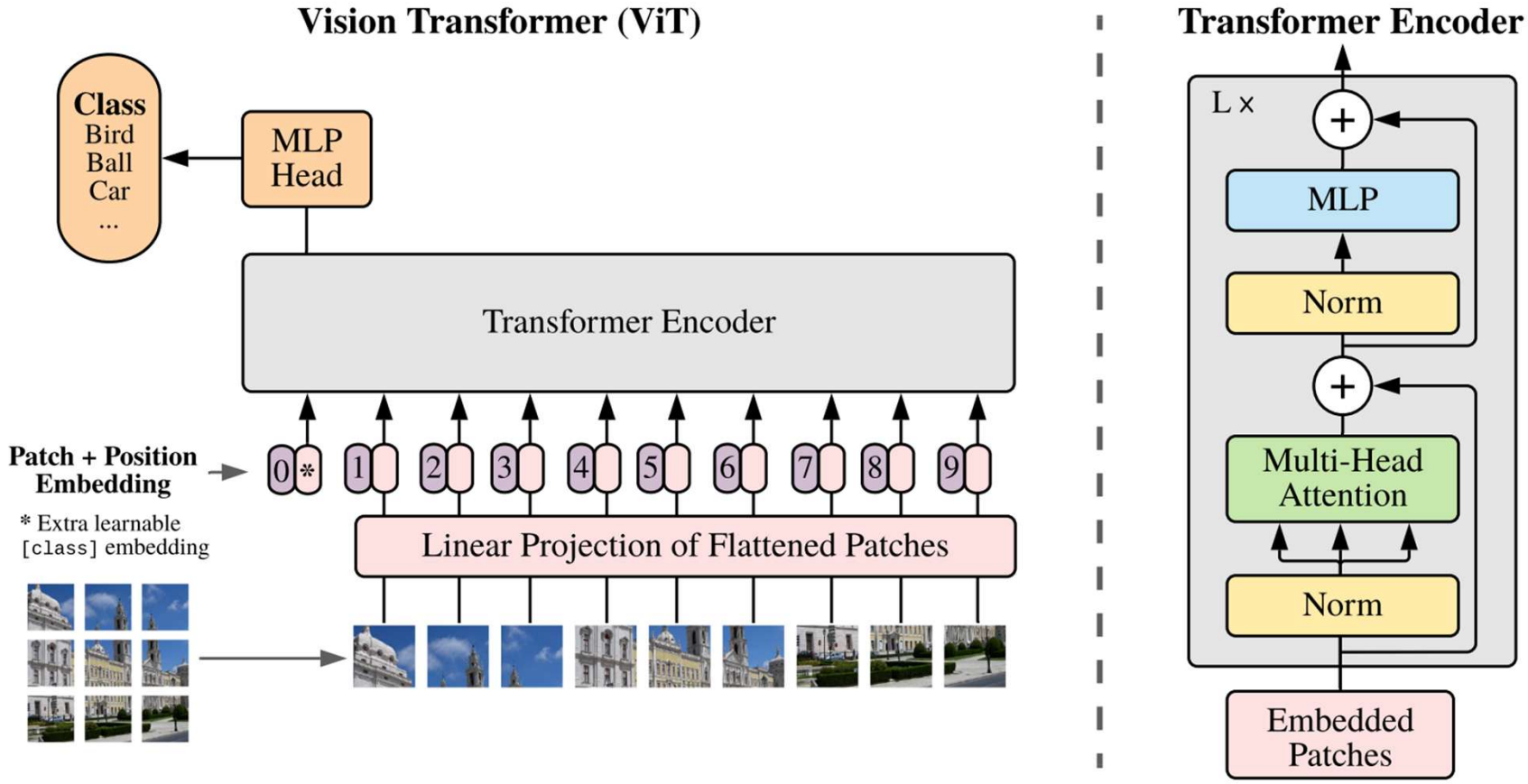
In ViT



In LLM

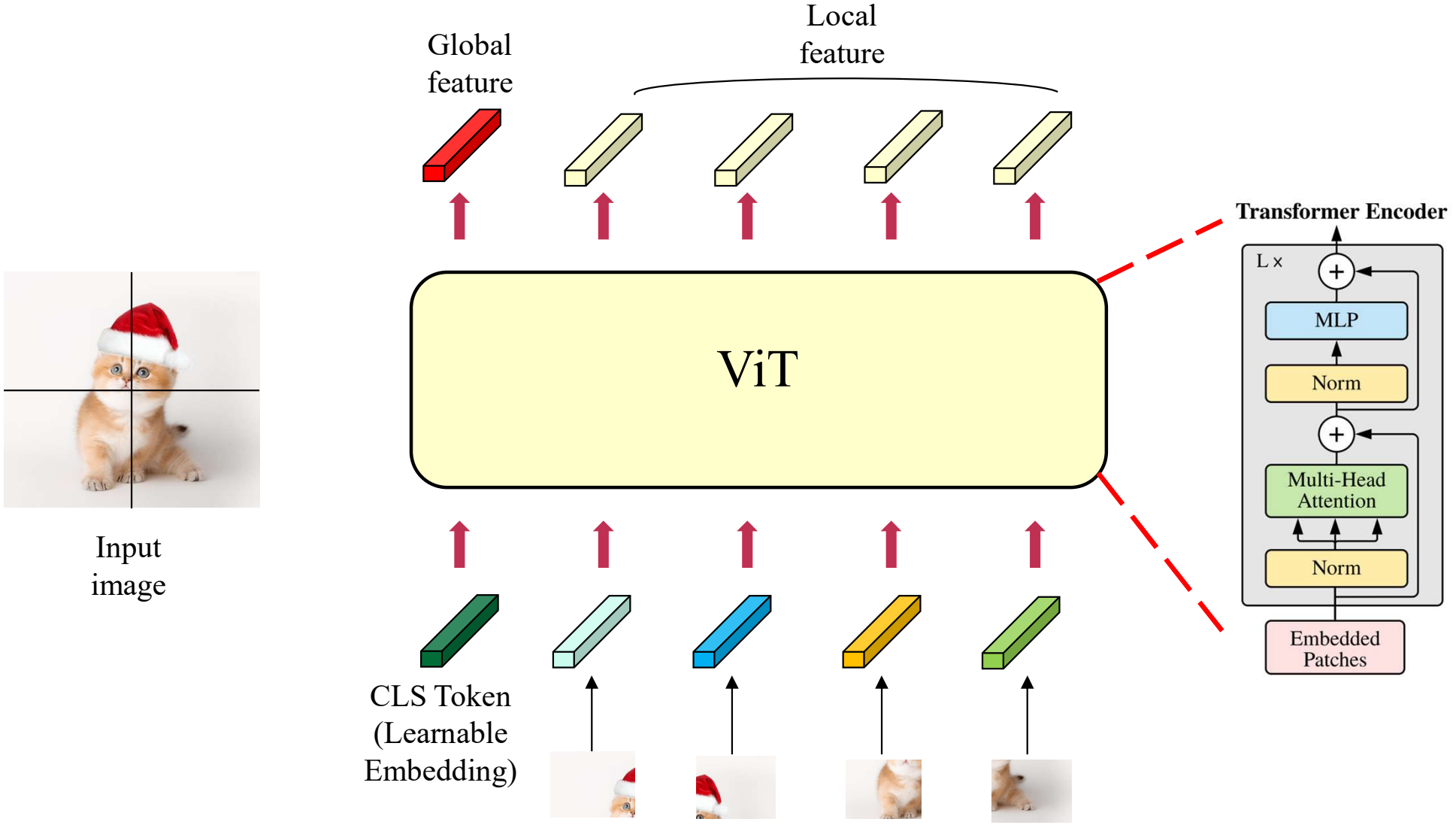
Background

- ViT Mechansim (Vision Modality)
  - 이미지를 고정 크기의 패치로 나눠 각 패치를 토큰처럼 처리하여 Transformer 구조로 학습하는 시각 인식 모델
    - Self-Attention을 통해 패치 간 관계를 학습하여 Visual feature 추출



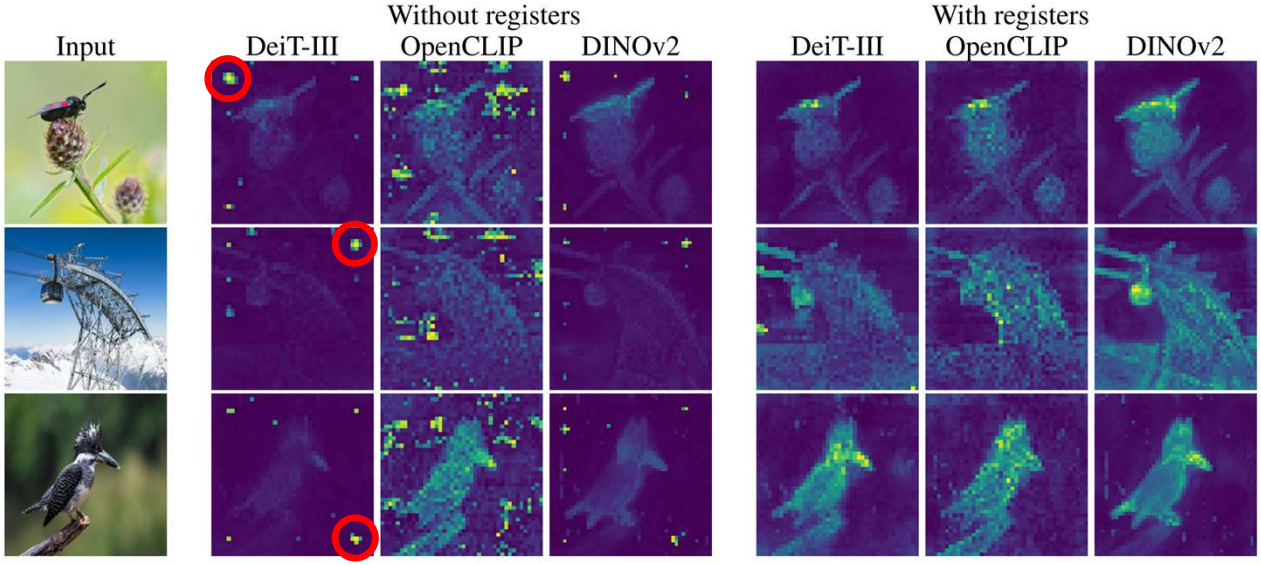
Background

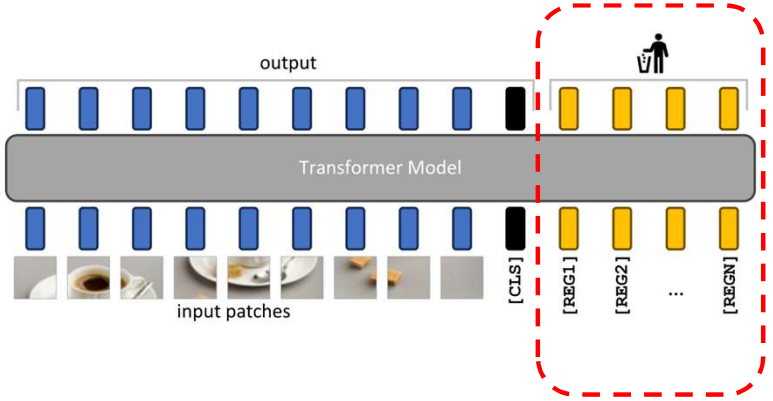
- ViT Mechansim (Vision Modality)
  - Detail



Paper I

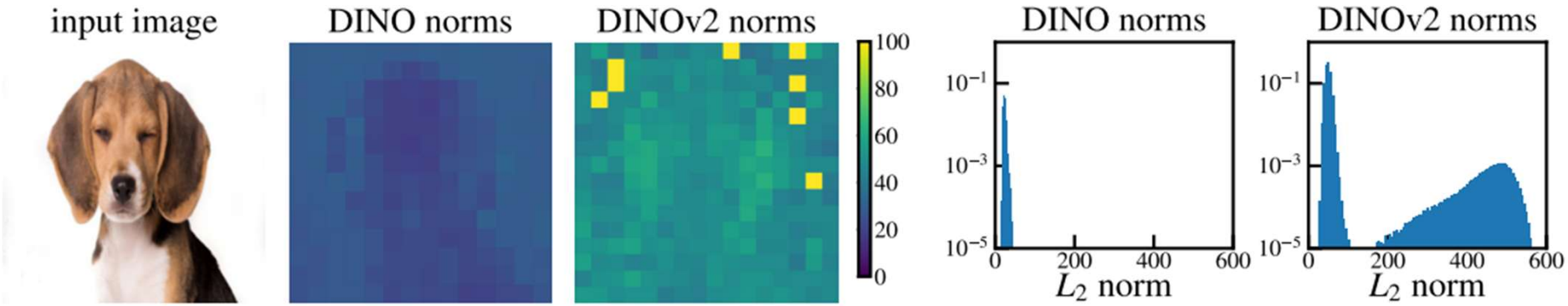
- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - ViT 모델에서 배경과 같은 의미 없는 토큰의 outlier 현상을 밝혀내고 분석
    - 주변 토큰들에 비해 약 10배 더 높은 norm 값을 가지고, 전체의 약 2%에 해당
    - 주로 중간 레이어에서 나타나고, 오래 학습하거나 모델이 큰 경우에 발생
    - ⇒**모델이 알아서 유용하지 않은 patch를 찾아 spatial 정보를 버리고 global 정보가 담기도록 학습한다고 주장**
  - 학습 때 register token을 추가하면 outlier 토큰들이 사라지고 Dense prediction 및 object discovery 성능이 향상된다고 주장





Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Problem Formulation
    - DINOv2 및 여러 최신 ViT(DeiT-3, OpenCLIP)에서 어텐션 맵에서 artifact를 보임
      - artifact: 출력 임베딩의 L2 norm이 매우 큰 Outlier 토큰
        - ✓논문에선 150 이상의 high-norm 토큰을 artifact로 정의함
      - *DINOv1에선 예외적으로 artifact 현상이 일어나지 않는다 함



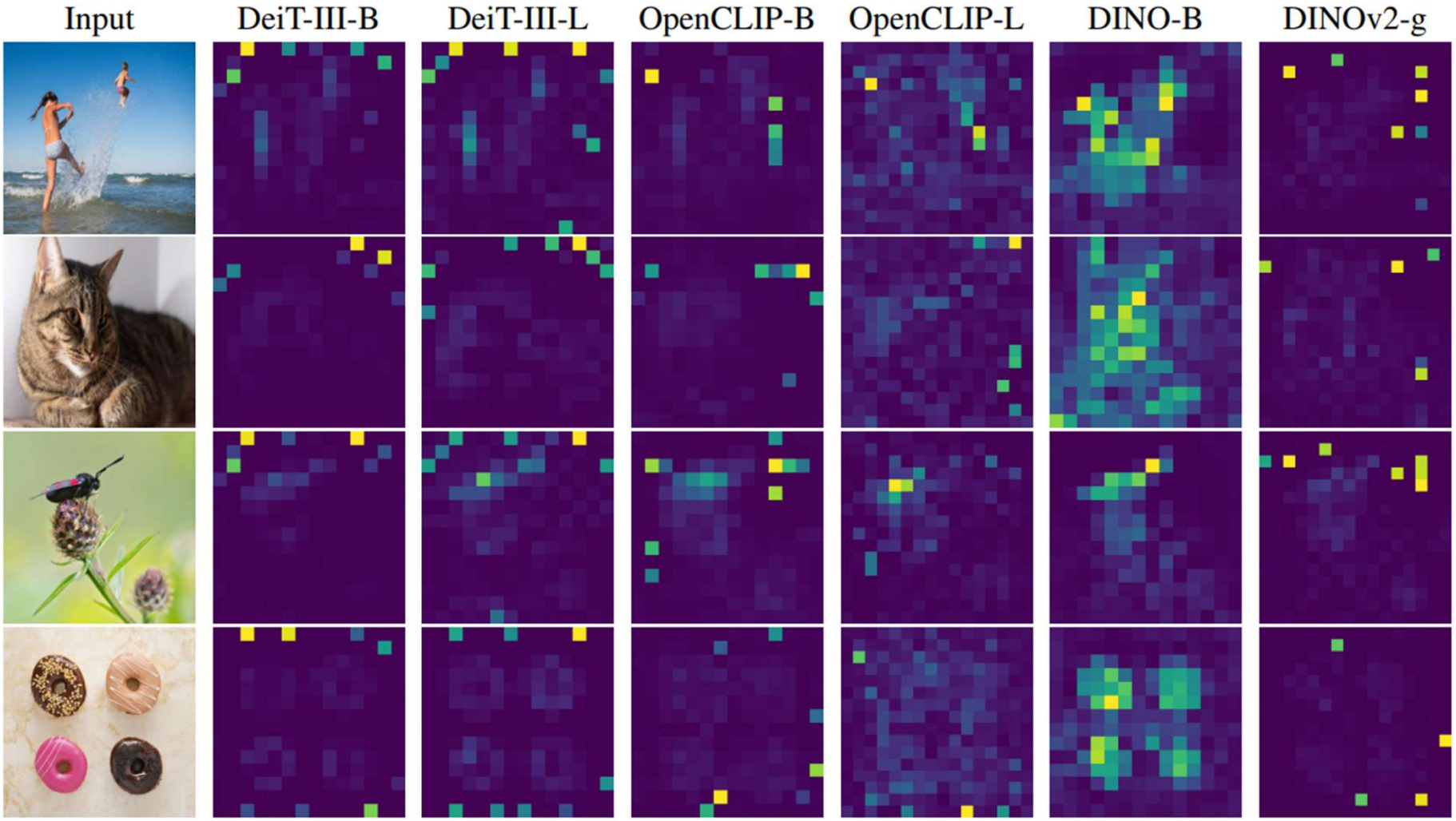
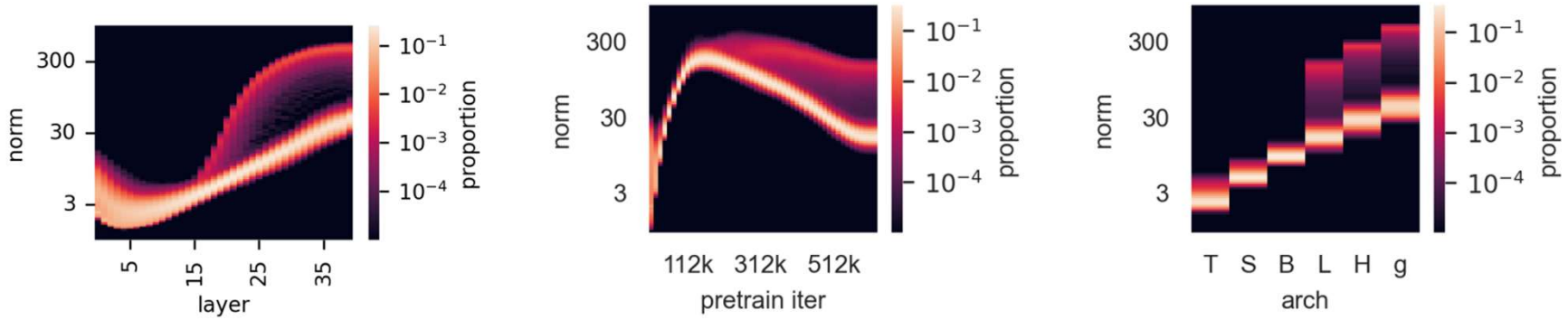


Figure 2: Illustration of artifacts observed in the attention maps of modern vision transformers. We consider ViTs trained with label supervision (DeiT-III), text-supervision (OpenCLIP) or self-supervision (DINO and DINOv2). Interestingly, all models but DINO exhibit peaky outlier values in the attention maps. The goal of this work is to understand and mitigate this phenomenon.

Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Problem Formulation
    - (a) 중반 레이어에서 부터 아티팩트 발생
    - (b) 전체 학습의 약 1/3 이후 부터 출현
    - (c) ViT-L 이상 모델에서만 발생

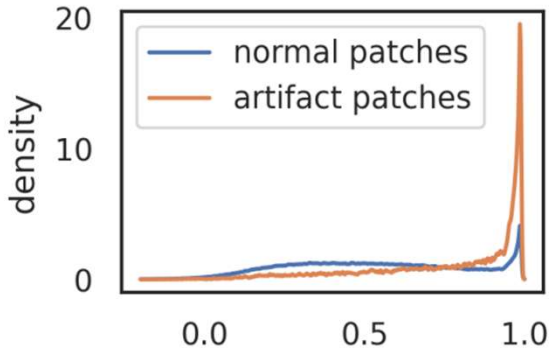


(a) Norms along layers. (b) Norms along iterations. (c) Norms across model size.

Figure 4: Illustration of several properties of outlier tokens in the 40-layer DINOv2 ViT-g model. **(a)**: Distribution of output token norms along layers. **(b)**: Distribution of norms along training iterations. **(c)**: Distribution of norms for different model sizes. The outliers appear around the middle of the model during training; they appear with models larger than and including ViT-Large.

Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Problem Formulation
    - **Artifact들은 배경과 같은 의미 없는 토큰에서 나타남**
      - Artifact 토큰들은 주변 토큰들과 유사한 경향을 보임, 중복도가 높음
    - **Artifact들은 Local information이 부족**
      - Position prediction 및 Reconstruction Task에서 normal 토큰들에 비해 성능이 낮음
        - ✓Position prediction: 해당 토큰의 위치를 맞추는 Task
          - Linear classifier를 학습하여 평가
        - ✓Pixel Reconstruction: 원래 패치의 RGB 픽셀값을 예측
          - Linear regression를 학습하여 평가



(a) Cosine similarity to neighbors.

	position prediction		reconstruction
	top-1 acc	avg. distance ↓	L2 error ↓
normal	**41.7**	**0.79**	**18.38**
outlier	22.8	5.09	25.23

(b) Linear probing for local information.

Paper I

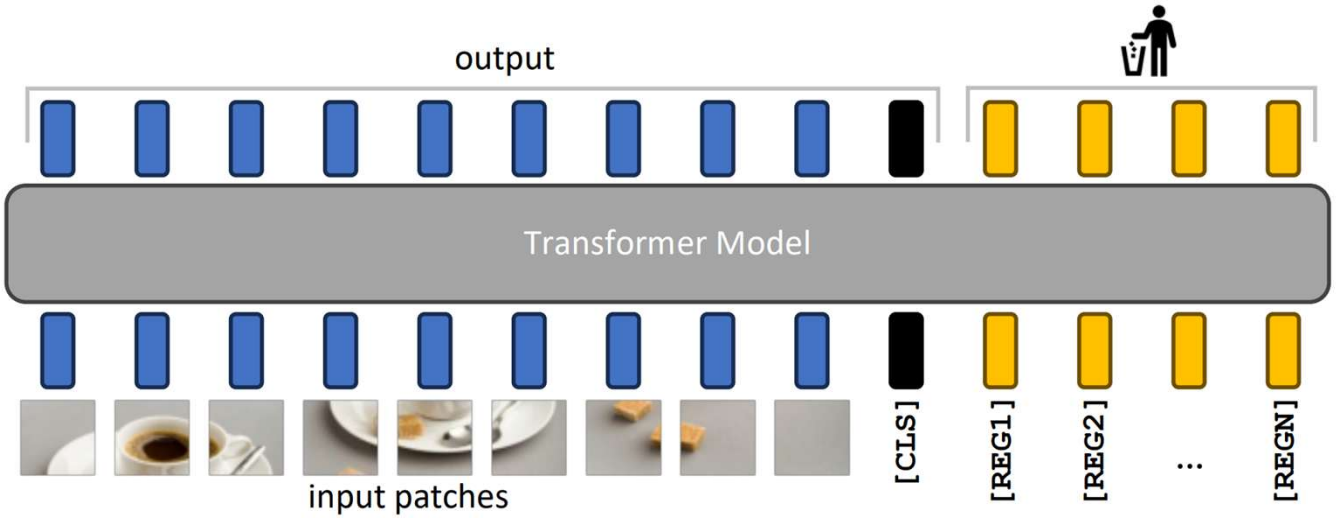
- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Problem Formulation
    - Artifact들은 Global information을 많이 보유함
      - Outlier token으로 classification 하였을 때 normal token들보다 성능이 좋음
        - ✓Caltech 101 Task의 경우 CLS Token보다 더 높은 성능을 냄

	IN1k	P205	Airc.	CF10	CF100	CUB	Cal101	Cars	DTD	Flow.	Food	Pets	SUN	VOC
[CLS]	**86.0**	**66.4**	**87.3**	**99.4**	**94.5**	**91.3**	96.9	**91.5**	**85.2**	**99.7**	**94.7**	**96.9**	**78.6**	89.1
normal	65.8	53.1	17.1	97.1	81.3	18.6	73.2	10.8	63.1	59.5	74.2	47.8	37.7	70.8
outlier	69.0	55.1	79.1	99.3	93.7	84.9	**97.6**	85.2	84.9	99.6	93.5	94.1	78.5	**89.7**

Table 1: Image classification via linear probing on normal and outlier patch tokens. We also report the accuracy of classifiers learnt on the class token. We see that outlier tokens have a much higher accuracy than regular ones, suggesting they are effectively storing global image information.

Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Hypothesis
    - Artifact의 발생 원인
      - 충분히 큰 규모의 모델이 충분히 학습되면, 모델은 중복된 토큰을 인식하고 이를 global information을 저장 처리 하는 용도로 활용
        - ✓중복된 토큰은 해당 패치의 local information을 잃어버려 Dense prediction과 같은 Local information이 중요한 Task에서 성능 저하가 일어 날 수 있음
        - ✓저자는 정확한 원인을 완전히 규명하지는 못했으나, 사전 학습 방식이 큰 영향을 미친다고 언급
    - 해결책
      - 모델이 자유롭게 global information을 저장 처리 할수 있는 새로운 토큰(Register)을 추가하여 학습
        - ✓학습 가능한 토큰이며, 해당 토큰을 학습 이후 Task에 활용하지 않고 폐기함



Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Experiments
    - Self-supervised, supervised 등 여러 방식으로 학습한 ViT 모델에 적용
      - DINOv2: unsupervised 방식으로 학습
      - OpenCLIP: Constrative(Self-supervised) 방식으로 학습
      - DeiT-3: Supervised 방식으로 학습
    - register 토큰을 추가하여 학습한 결과 artifact들이 효과적으로 감소함
      - 한 토큰 만으로도 눈에 뛰게 감소

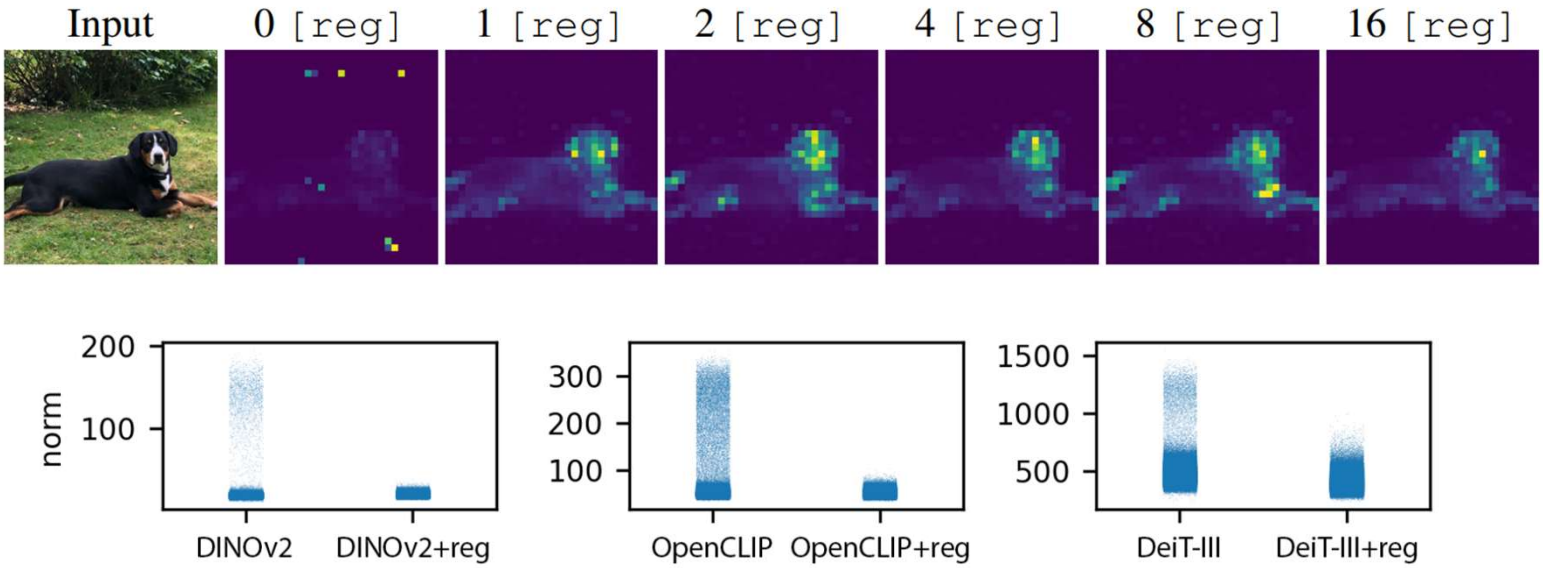


Figure 7: Effect of register tokens on the distribution of output norms on DINOv2, OpenCLIP and DeiT-III. Using register tokens effectively removes the norm outliers that were present previously.

Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Experiments
    - 성능 결과 분석
      - imageNet(classification): register를 추가한 후에도 global information이 중요한 task에서 성능 저하가 크지 않음
      - ADE20K(segmentation): 성능 소폭 상승
      - NYUd(Dense estimation): 성능 소폭 상승

	ImageNet Top-1	ADE20k mIoU	NYUd rmse ↓
DeiT-III	84.7	38.9	0.511
DeiT-III+reg	84.7	39.1	0.512
OpenCLIP	78.2	26.6	0.702
OpenCLIP+reg	78.1	26.7	0.661
DINOv2	84.3	46.6	0.378
DINOv2+reg	84.8	47.9	0.366

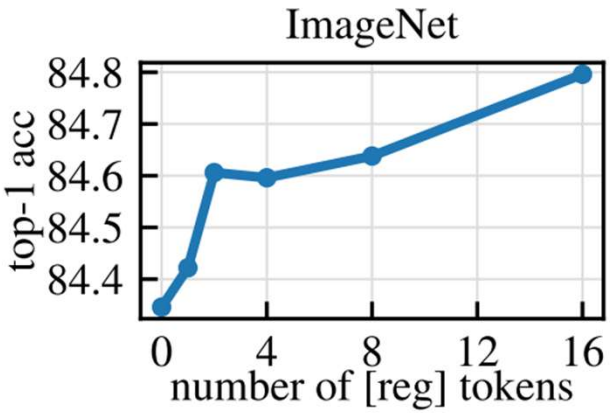
(a) Linear evaluation with frozen features.

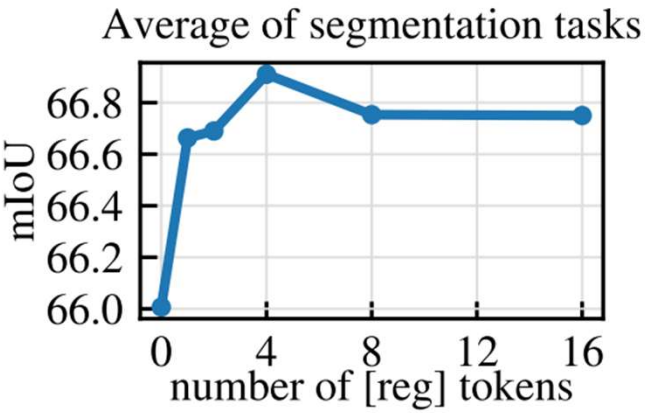
	ImageNet Top-1
OpenCLIP	59.9
OpenCLIP+reg	60.1

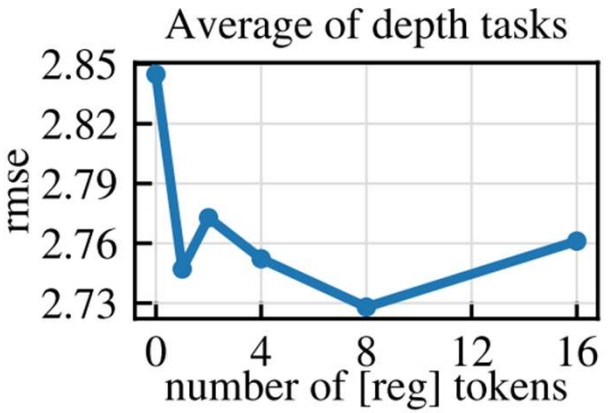
(b) Zero-shot classification.

Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Experiments
    - Number of register token







Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Experiments
    - Register 토큰의 어텐션 맵을 살펴보면, 객체의 서로 다른 부분에 주목하는 흥미로운 경향을 보임
      - 이는 강제로 설계된 것이 아니라, 학습 과정에서 다양성이 자연스럽게 형성됨

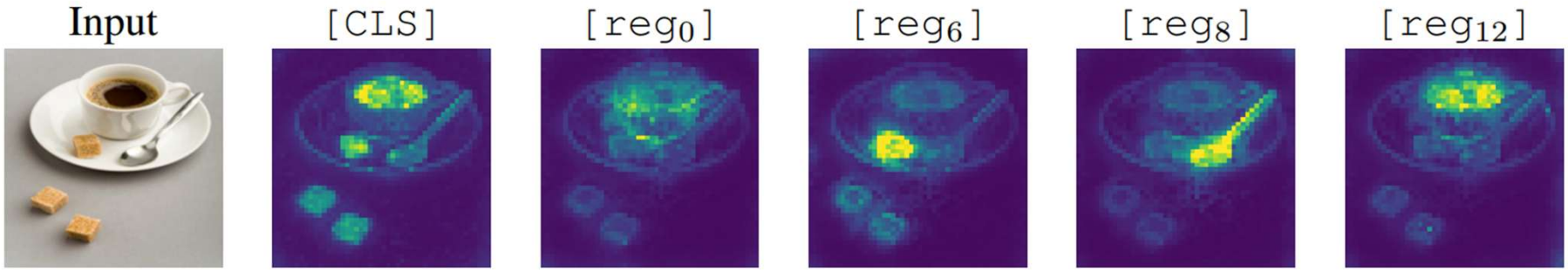


Figure 9: Comparison of the attention maps of the [CLS] and register tokens. Register tokens sometimes attend to different parts of the feature map, similarly to slot attention (Locatello et al., 2020). This behaviour was never required from the model, and emerged naturally from training.

Paper I

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Conclusion
    - 본 연구는 대형 ViT 모델에서 배경과 같이 비교적 유용하지 않은 토큰에서 발생하는 아티팩트 현상을 규명·분석함
    - Register 토큰을 추가해 이러한 아티팩트를 제거하는 간단한 방법으로 제안한 가설을 검증하였다.

Background

- LLM Mechansim (Language Modality)
  - 주로 트랜스포머의 디코더로만 구성 (Decoder Only)
    - Autoregressive 방식으로 다음 단어를 차례대로 예측

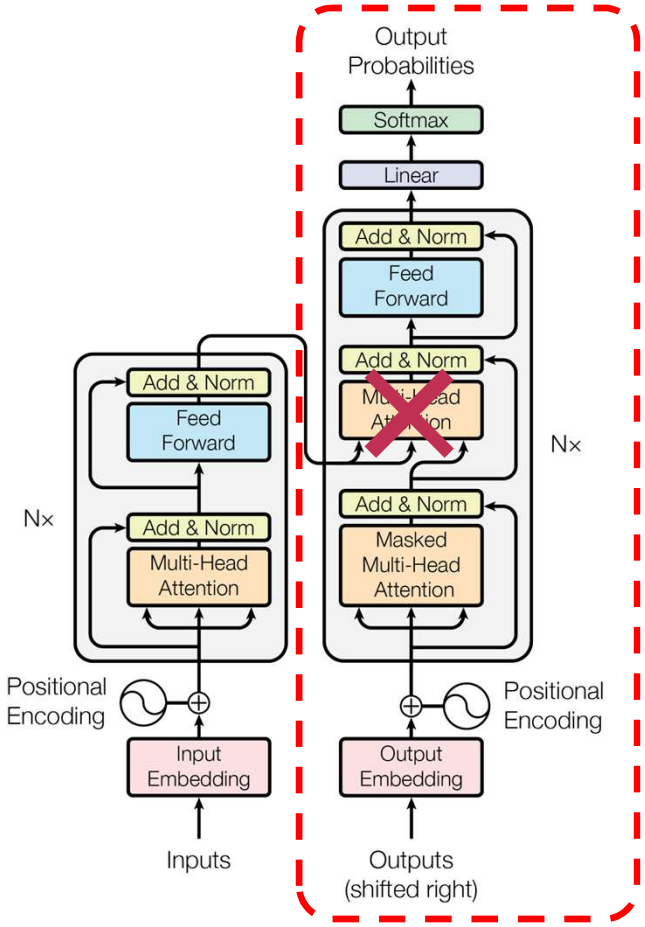
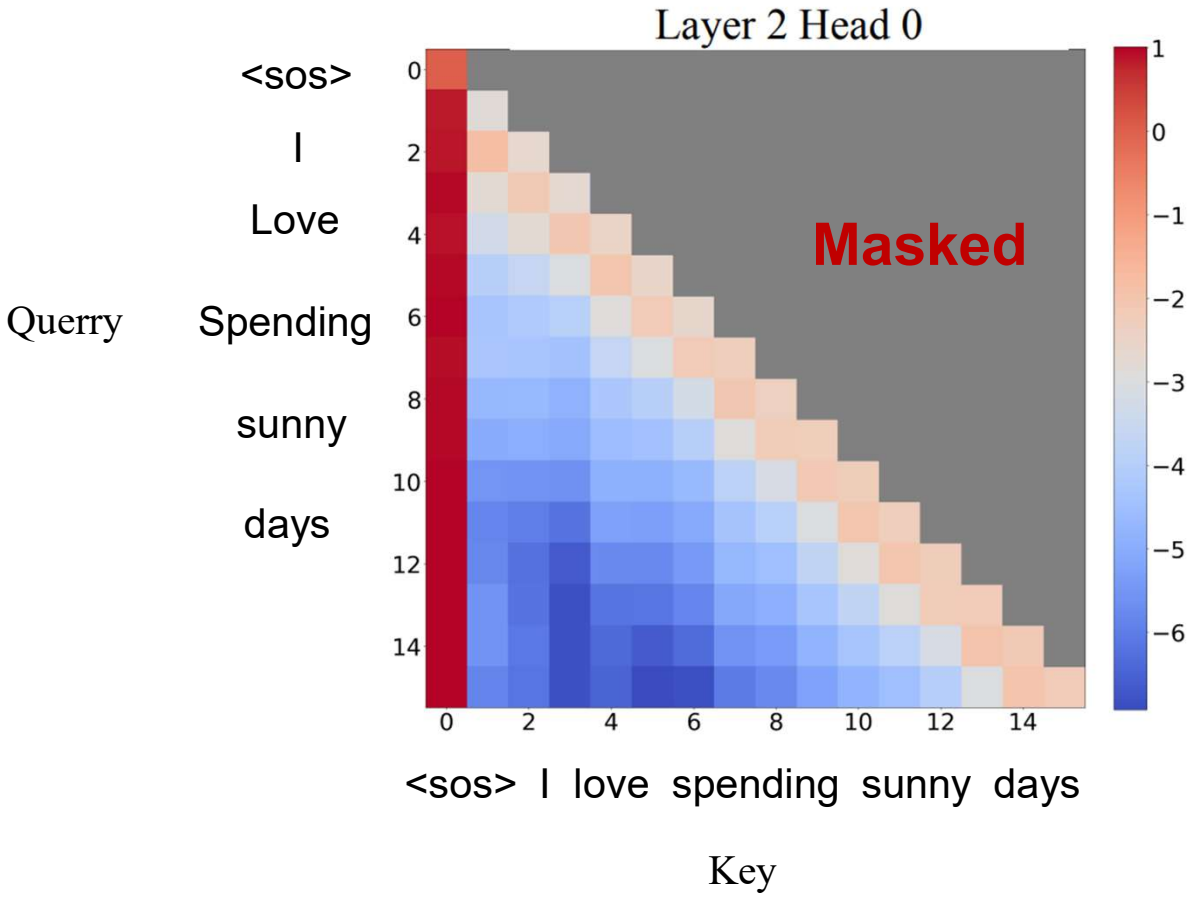


Figure 1: The Transformer - model architecture.

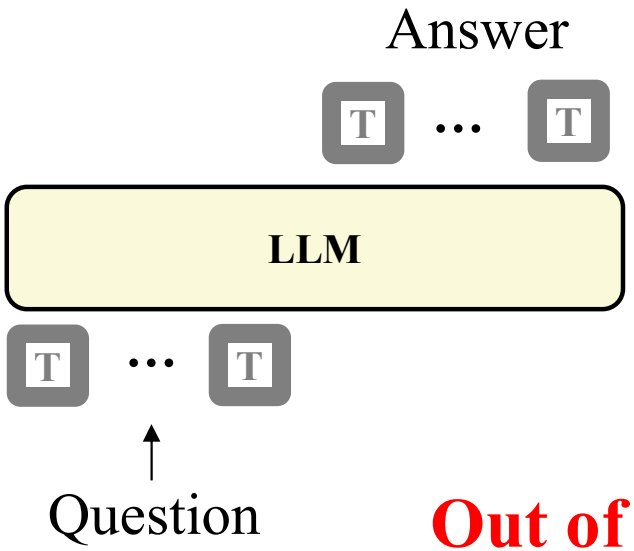
Background

- LLM Mechansim (Language Modality)
  - LLM attention map



Background

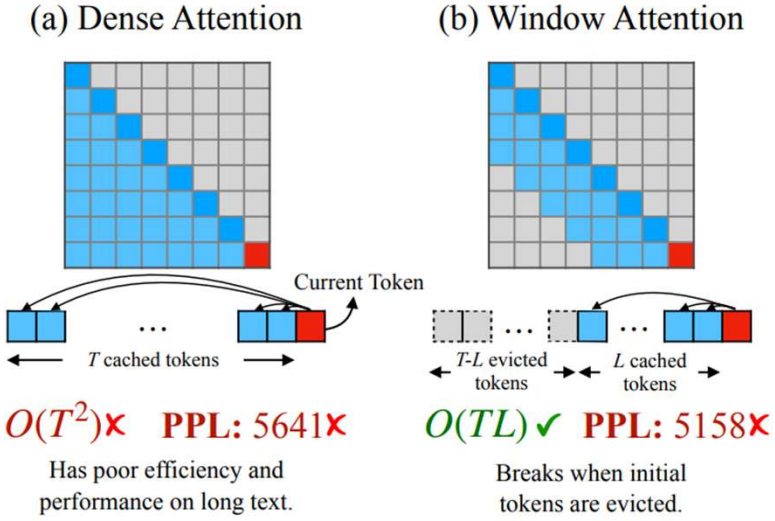
- LLM Mechansim (Language Modality)
  - KV Cache (Key-Value Cache)
    - Transformer의 입력 길이가 길어질 수록 연산 비용 및 메모리 사용량 증가
      - Attention 연산 복잡도는 $O(n^2)$에 비례
    - 매번 Key-Value를 재계산하지 않고 KV Cache를 저장하여 반복적인 Attention 연산을 피함
      - KV-Cache 활용 시 $O(n)$까지 감소 가능
      - 그러나 KV Cache가 누적되면 메모리 사용량이 지속적으로 증가하는 문제가 존재



Auto regressive Decoding

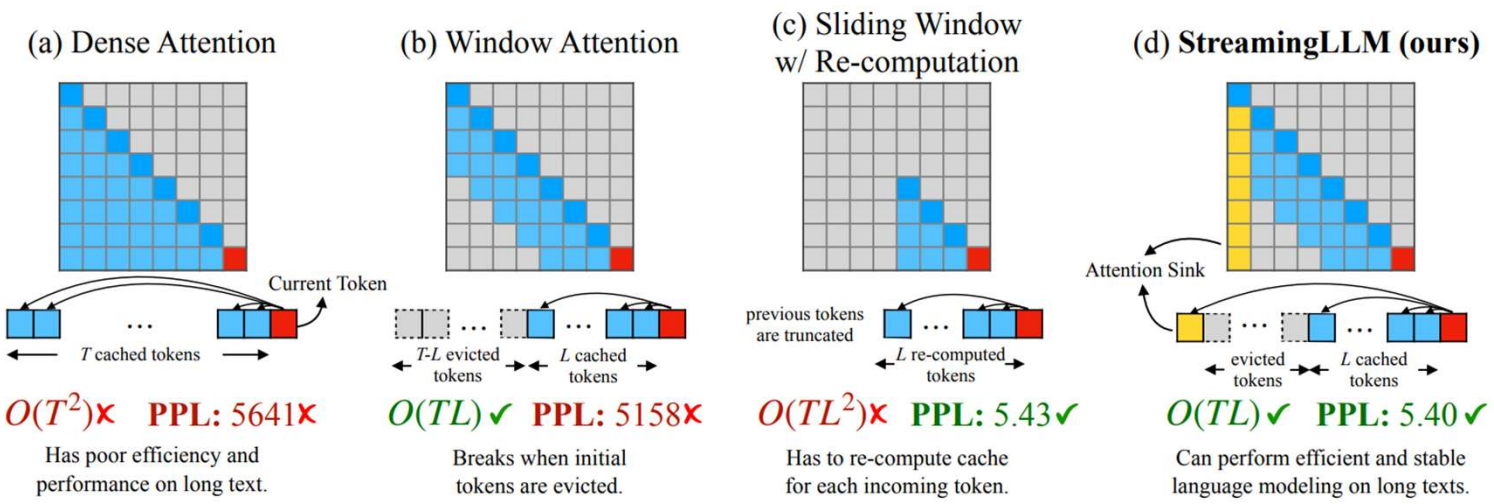
1 + 4 + 9 + … +10000

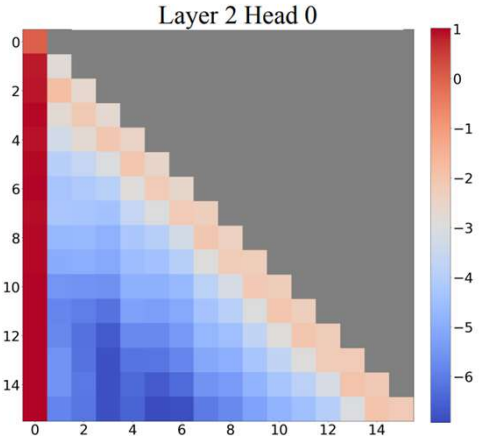
Out of Memory



Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - LLM 내부 어텐션에서 맨 앞 토큰에 높은 어텐션이 할당되는 현상(attention sink)을 규명·분석
    - LLM의 KV cache에서 해당 토큰(sink token)이 제거될 경우 성능이 급격히 저하되어, sink 토큰 보존의 중요성을 확인



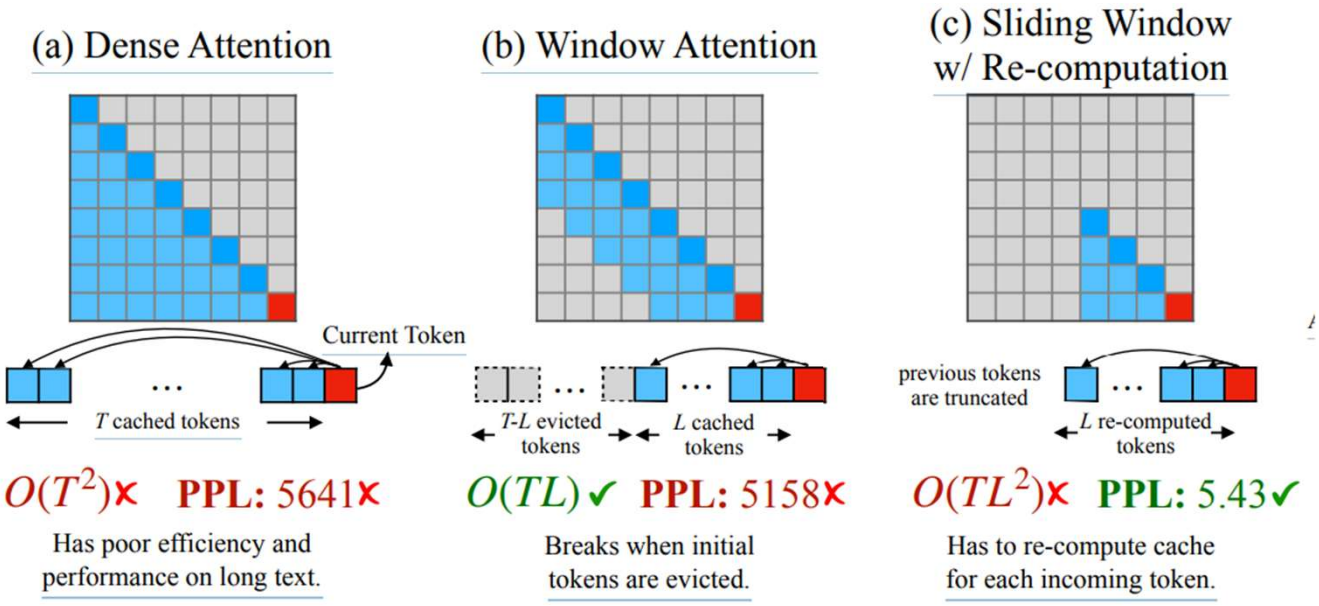


Perplexity (PPL): 언어 모델에 주로 쓰이는 지표로 모델 응답의 entrophy와 관련됨, 낮을수록 좋음

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Introduction
    - 1. LLM은 긴 시퀀스(context)를 처리할 때 KV 캐시가 계속 쌓여 메모리 사용량과 latency가 증가하는 문제가 있음
    - 2. 연산 문제뿐만 아니라 사전 학습된 윈도우 길이를 넘어서면 성능이 급격히 저하되는 문제가 있음

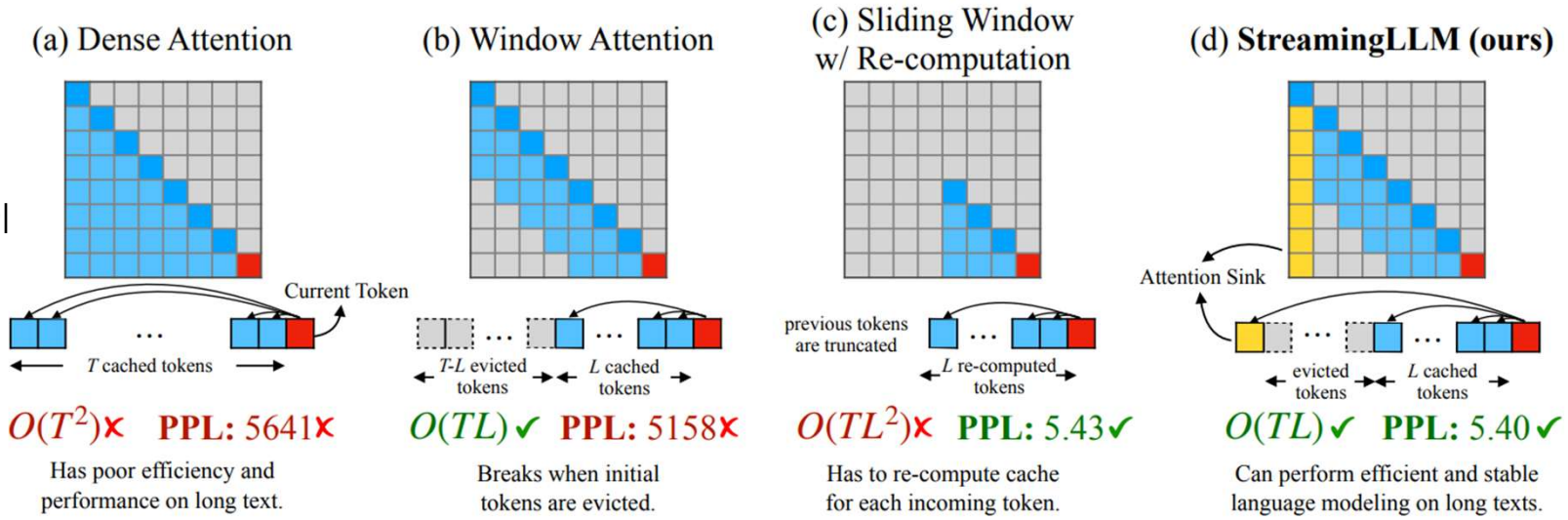
    => **"효율성과 성능을 희생하지 않고, 무한 길이 입력에 대해 LLM을 deploy 할 수 있는가?"**



Paper II

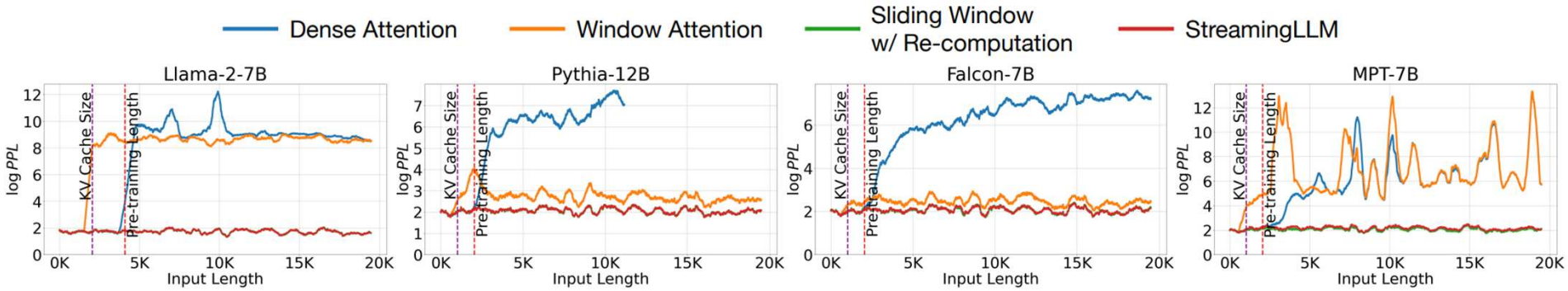
- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Introduction
    - (a) Dense Attention: 앞에 있던 모든 단어를 KV Cache에 저장
      - 효율성 저하 및 사전 윈도우 길이 초과
    - (b) Window Attention: 최근 L개의 KV만 Caching
      - 효율성은 좋지만 초기 토큰이 윈도우에서 제거되었을 때 급격한 성능 저하를 보임
    - (c) 최근 L 토큰의 KV를 매번 재계산
      - 성능은 유지되지만 연산량 급증
    - (d) 초기 토큰 및 최근 L개의 KV만 Caching
      - 효율적이면서도 성능을 유지함

T: 전체 토큰 개수
L: 사전 윈도우 길이

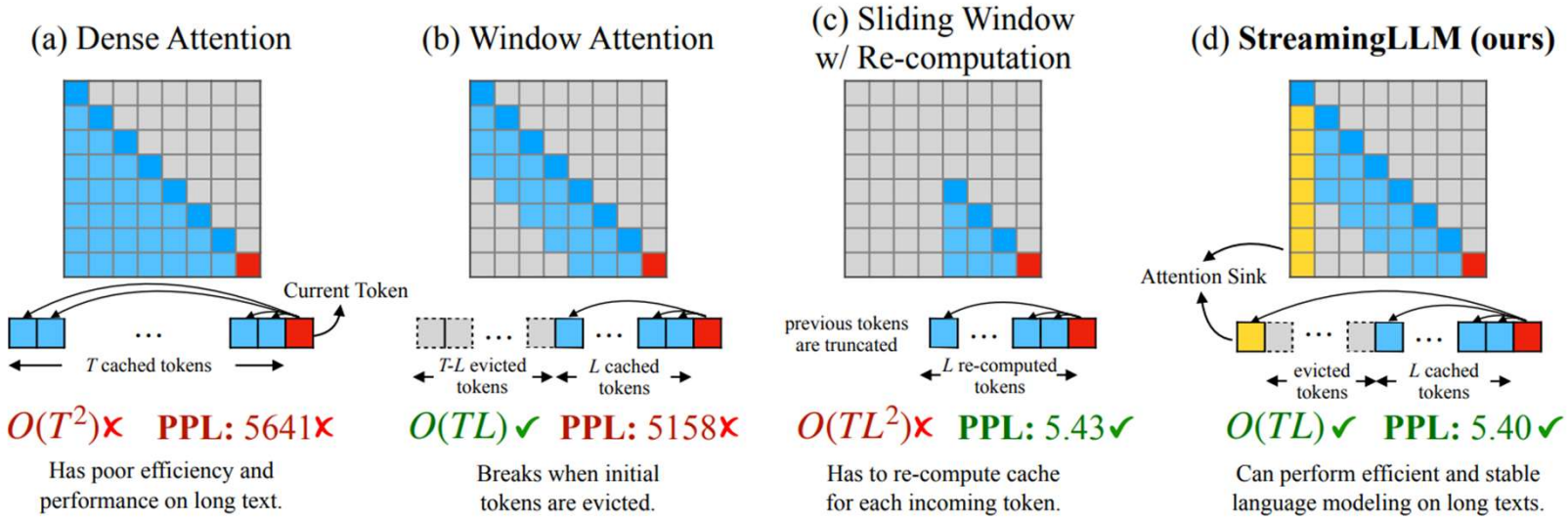


Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Introduction
    - KV Cache size 2048, Pre-training length 4096



T: 전체 토큰 개수
L: 사전 윈도우 길이



Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Introduction
    - 왜 초기 토큰이 Cache에서 제거되면 급격한 성능 저하를 보이는가?
      - LLM 내부 layer 2 이후 모든 레이어 및 헤드에서 text token들이 초기 토큰에 어텐션을 과도하게 집중하는 현상을 발견함 **(attention sink)**
        - ✓초기 토큰을 제거하게 되면 어텐션 계산시 Softmax 분모의 상당 부분이 사라지게 되어 어텐션 스코어 분포가 급격히 변하게 됨

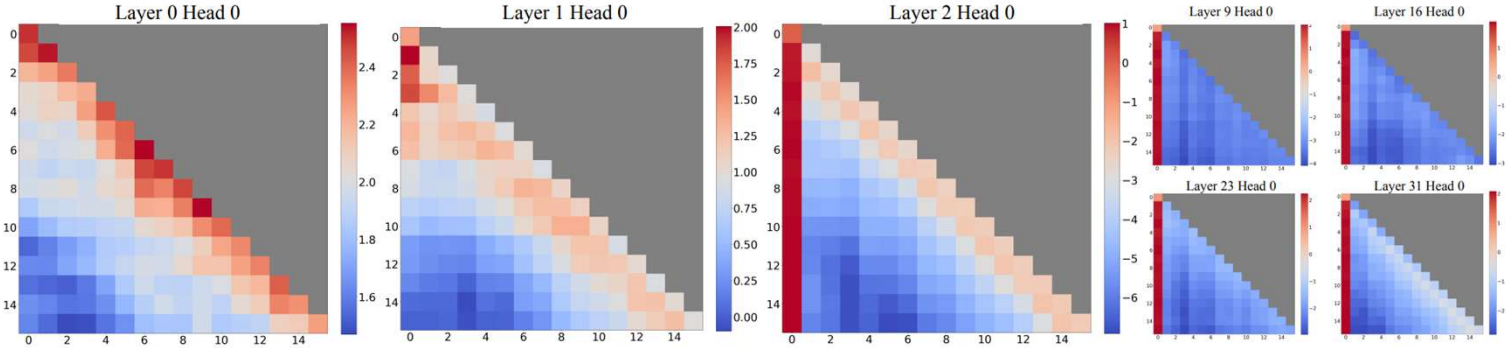


Figure 2: Visualization of the *average* attention logits in Llama-2-7B over 256 sentences, each with a length of 16. Observations include: (1) The attention maps in the first two layers (layers 0 and 1) exhibit the "local" pattern, with recent tokens receiving more attention. (2) Beyond the bottom two layers, the model heavily attends to the initial token across all layers and heads.

$$\text{SoftMax}(x)_i = \frac{e^{x_i}}{e^{x_1} + \sum_{j=2}^{N} e^{x_j}},$$

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Method
    - 왜 초기 토큰에게 과도하게 어텐션을 집중하는가?
      - 저자는 두 가지 가능성에 대해 실험을 진행
        - ✓1. 초기 토큰 "<sos>"의 의미가 중요하기 때문에 집중한다
        - ✓**2. 의미와 상관없이 토큰이 초기에 있다는 절대적인 위치때문에 집중한다**
      - 초기 토큰 (<sos>) 위치에 줄바꿈 토큰(\n)을 대체하여도 attention sink 현상이 발생함
        - ✓초기 토큰의 의미보단 절대적인 위치가 중요함
          - 저자는 언어 모델의 Autoregressive 특성 때문에 초기 토큰들이 모든 어텐션에 참여하므로 불필요한 어텐션을 흡수하는 sink 역할로 더 쉽게 학습된다고 설명함
          - 모델은 어텐션 값을 반드시 분배해야 하므로 불필요한 어텐션 값을 특정 토큰에 몰아넣는 경향이 존재함

Llama-2-13B	PPL (↓)
0 + 1024 (Window)	5158.07
4 + 1020	5.40
4"\n"+1020	5.60

$$\text{SoftMax}(x)_i = \frac{e^{x_i}}{e^{x_1} + \sum_{j=2}^{N} e^{x_j}},$$

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Method
    - Llama-2-7b는 초기 4개의 토큰까지 어텐션 싱크로 활용
      - 사전 학습 시 모든 입력 샘플의 시작에 동일한 시작 토큰이 없었기 때문
        - ✓각 문단 앞에 <s> 토큰을 붙이지만, 텍스트를 Chunk로 나누는 과정 때문에 실제 0번째 위치에는 랜덤한 토큰이 오는 경우가 많음
          - 일관성 없는 시작 토큰 때문에 모델은 여러 개의 초기 토큰을 싱크로 사용

<s> One notable example is the formation of high-norm outlier tokens, which tend to appear only after a significant portion of training and primarily in deeper layers. These tokens often correspond / **to** redundant or low-information patches, yet they attract disproportionate attention from the model. Understanding when and why such artifacts emerge is crucial for improving interpretability and optimizing the efficiency of vision

Cache Config	0+2048	1+2047	2+2046	4+2044	8+2040
Falcon-7B	17.90	12.12	12.12	12.12	12.12
MPT-7B	460.29	14.99	15.00	14.99	14.98
Pythia-12B	21.62	11.95	12.09	12.09	12.02
Cache Config	0+4096	1+4095	2+4094	4+4092	8+4088
Llama-2-7B	3359.95	11.88	10.51	9.59	9.54

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Method
    - Learnble Sink
      - LLM 모델의 0번째 위치에 새로 학습가능한 토큰을 넣어 하나의 토큰만으로 Sink 역할을 하게함
        - ✓Vision need register의 register와 비슷한 역할 수행
        - ✓추가 후 윈도우에 하나의 sink 토큰만 추가하더라도 성능이 유지됨
      - Zero Sink: 변형 Softmax 수식을 활용하여 어텐션 분포의 변화를 완화
        - ✓Learnable Sink Token 방식만큼의 안정성을 제공하지는 못함

      => 저자들은 LLM 모델을 배포할때 Sink 토큰을 추가하여 학습 하는것을 권장함

Cache Config	0+1024	1+1023	2+1022	4+1020
Vanilla	27.87	18.49	18.05	18.05
Zero Sink	29214	19.90	18.27	18.01
Learnable Sink	1235	**18.01**	18.01	18.02

$$\text{SoftMax}_1(x)_i = \frac{e^{x_i}}{1 + \sum_{j=1}^{N} e^{x_j}},$$

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Experiments
    - Model
      - Llama-2
      - MPT
      - Pythia
      - Falcon
    - Input length가 길어지더라도 PPL이 급격히 증가하지 않음

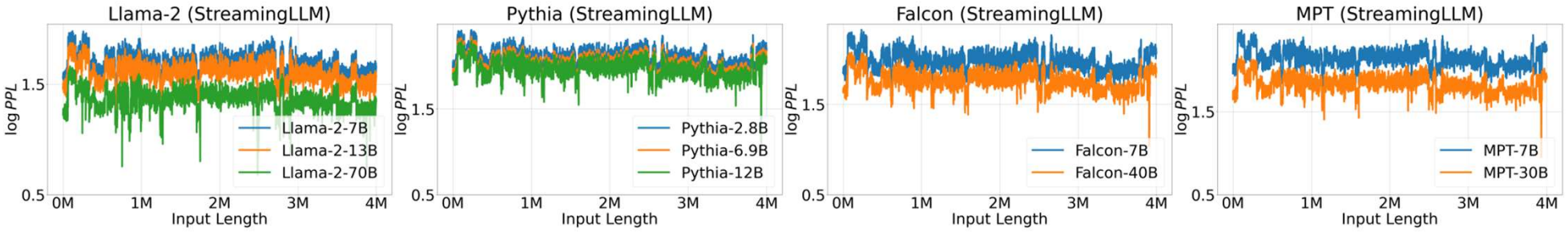
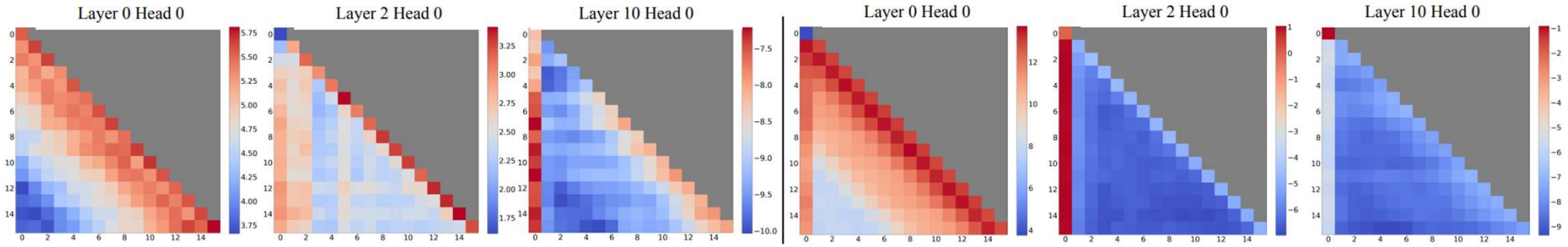


Figure 5: Language modeling perplexity of StreamingLLM on super long texts with 4 million tokens across various LLM families and scales. The perplexity remains stable throughout. We use the concatenated test set of PG19 (100 books) to perform language modeling, with perplexity fluctuations due to book transitions.

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Experiments
    - w/o sink token vs w/ sink token
      - Sink token 추가하여 학습한 경우 소폭의 성능 향상을 보임



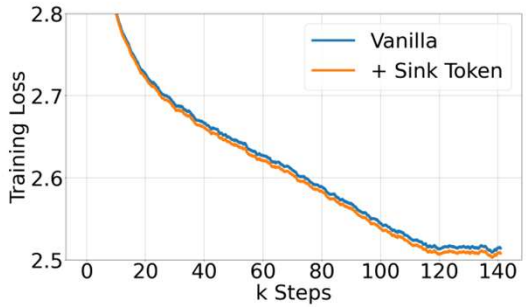


Figure 6: Pre-training loss curves of models w/ and w/o sink tokens. Two models have a similar convergence trend.

Table 4: Zero-shot accuracy (in %) across 7 NLP benchmarks, including ARC-[Challenge, Easy], HellaSwag, LAMBADA, OpenbookQA, PIQA, and Winogrande. The inclusion of a sink token during pre-training doesn't harm the model performance.

Methods	ARC-c	ARC-e	HS	LBD	OBQA	PIQA	WG
Vanilla	18.6	45.2	29.4	39.6	16.0	62.2	50.1
+Sink Token	**19.6**	**45.6**	**29.8**	**39.9**	**16.6**	**62.6**	**50.8**

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Experiments
    - Cache Sizes
      - 캐시 크기를 계속 늘려도 PPL이 일관되게 감소하지 않음
        - ✓Pre-training length = 4096

Cache	4+252	4+508	4+1020	4+2044
Falcon-7B	13.61	12.84	**12.34**	12.84
MPT-7B	**14.12**	14.25	14.33	14.99
Pythia-12B	13.17	12.52	**12.08**	12.09

Cache	4+508	4+1020	4+2044	4+4092
Llama-2-7B	9.73	9.32	**9.08**	9.59

Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Experiments
    - Efficiency Results
      - 재계산 기반 슬라이딩 윈도우 방식에 비해 토큰당 최대 22.2배의 속도 향상을 달성
        - ✓Single A6000 측정

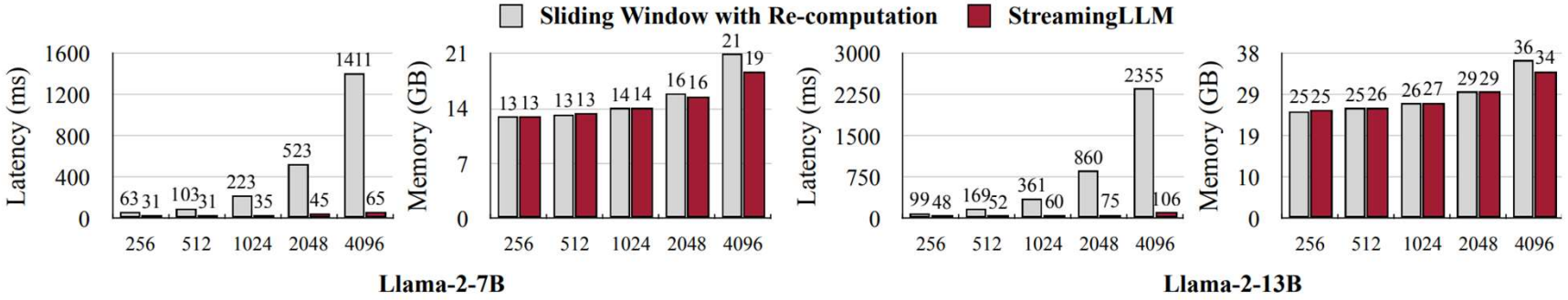
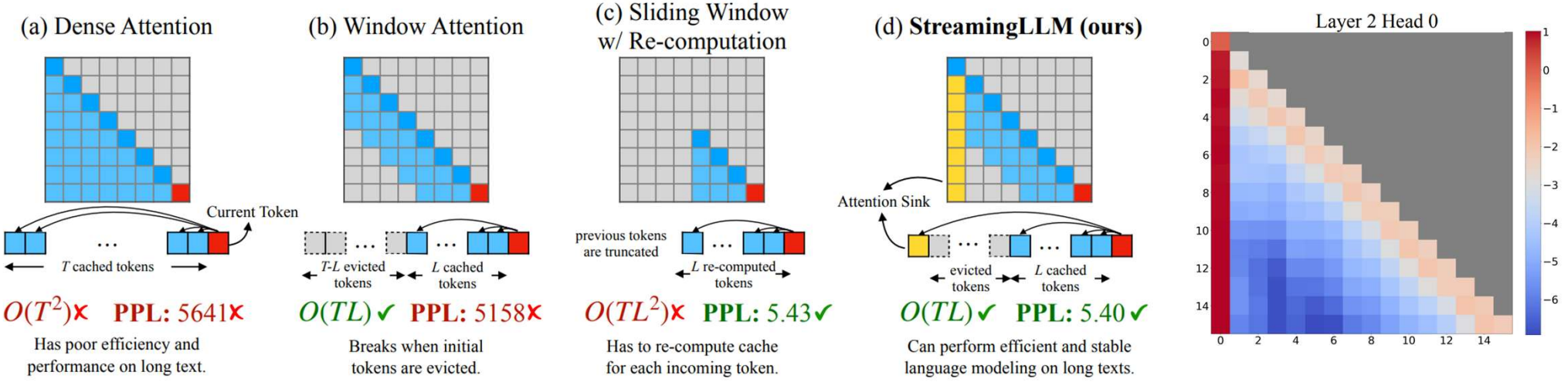


Figure 10: Comparison of per-token decoding latency and memory usage between the sliding window approach with re-computation baseline and StreamingLLM, plotted against the cache size (attention window size) on the X-axis. StreamingLLM delivers a remarkable speedup of up to 22.2× per token and retains a memory footprint similar to the re-computation baseline.

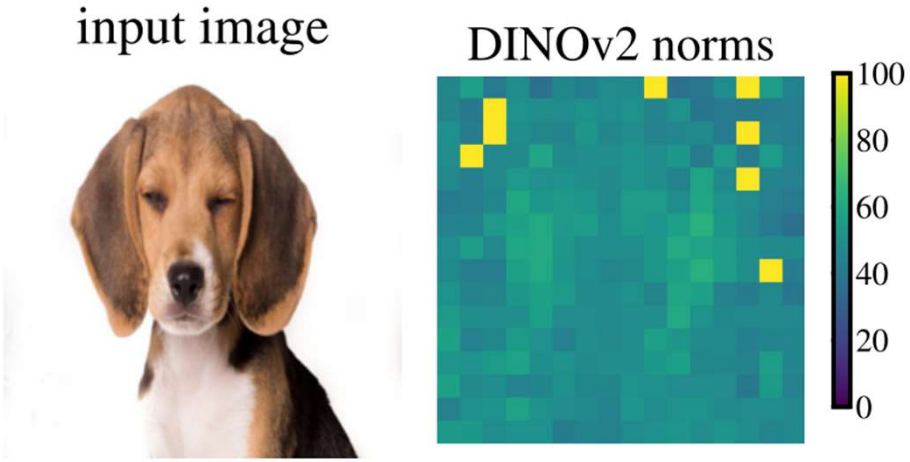
Paper II

- EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS (ICLR 2024)
  - Conclusion
    - LLM에서의 긴 시퀀스는 효율성 저하 및 성능 저하를 초래함
      - 윈도우 어텐션은 길이 문제는 해결하였지만 초기 토큰이 제외되어 성능이 급격히 하락하는 한계가 존재
    - 저자는 초기 토큰이 attention sink 역할을 규명하고, 이를 활용한 간단한 방법을 제안하여 기존 윈도우 어텐션의 문제를 해결했다.

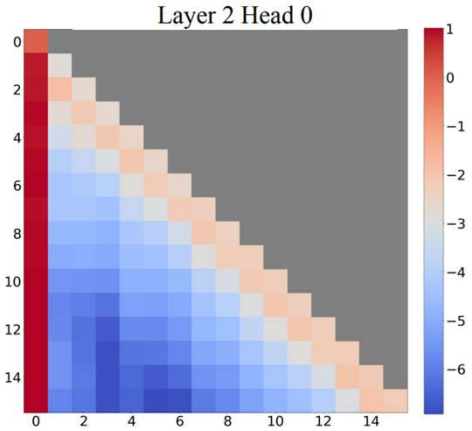


Conclusion

- Paper I vs Paper II
  - 대형 트랜스포머 구조 전반에서 어텐션이 과도하게 집중되는 현상이 관찰됨
    - ViT에선 배경과 같은 의미 없는 토큰에 집중됨
    - LLM에선 Autoregressive한 특성 때문에 초기 토큰에 집중됨
    - ⇒ 이는 모달리티를 구분하지 않고 공통적으로 나타나는 특성으로, 트랜스포머의 학습 과정에서 안정적인 예측을 위해 일어나는 현상이라 생각됨
    - ⇒ 트랜스포머 구조의 해석가능성 연구에 중요한 시사점을 제공



In ViT



In LLM

Supplements

- VISION TRANSFORMERS NEED REGISTERS (ICLR 2024 Oral)
  - Experiments
    - Object Discovey Task
      - LOST 알고리즘을 활용하였을 때 register token 추가 후 높은 성능 향상
        - ✓OpenCLIP에선 성능 저하가 일어남
          - 저자는 LOST 실험에서 OpenCLIP의 특징을 직접 사용하는 것이 아니라, 어텐션 맵 계산에서의 values를 사용하기 때문이라고 언급
          - Value에선 artifact가 잘 발생하지 않기 때문에 성능 향상이 없었다고 언급함

	VOC 2007	VOC 2012	COCO 20k
DeiT-III	11.7	13.1	10.7
DeiT-III+reg	27.1	32.7	25.1
OpenCLIP	38.8	44.3	31.0
OpenCLIP+reg	37.1	42.0	27.9
DINOv2	35.3	40.2	26.9
DINOv2+reg	55.4	60.0	42.0

Table 3: Unsupervised Object Discovery using LOST (Siméoni et al., 2021) on models with and without registers. We evaluated three types of models trained with various amounts of supervision on VOC 2007, 2012 and COCO. We measure performance using corloc. We observe that adding register tokens makes all models significantly more viable for usage in object discovery.