Massive Activations in Large Language Models

Mingjie Sun[1] Xinlei Chen[2] J. Zico Kolter[1,3] Zhuang Liu[2]
[1]Carnegie Mellon University [2]Meta AI Research [3]Bosch Center for AI

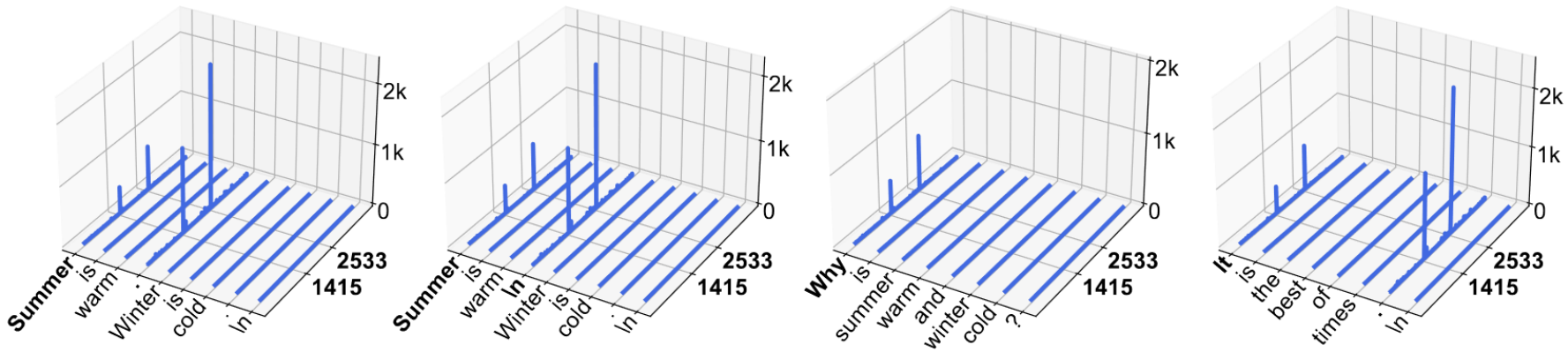


Figure 1: **Activation Magnitudes (z-axis) in LLaMA2-7B.** **x** and **y** axes are sequence and feature dimensions. For this specific model, we observe that activations with massive magnitudes appear in two fixed feature dimensions (1415, 2533), and two types of tokens—the starting token, and the first period (`.`) or newline token (`\n`).

Zou Lexiao

2024/03/25

Massive Activations

- Definition:

  Examining the hidden states in these models, we find that certain activations exhibit huge magnitudes, e.g., ***more than 4 orders of magnitude*** *larger than the median, and could take on absolute values larger than 15,000 in LLaMA2-70B, despite the presence of normalization layers.*

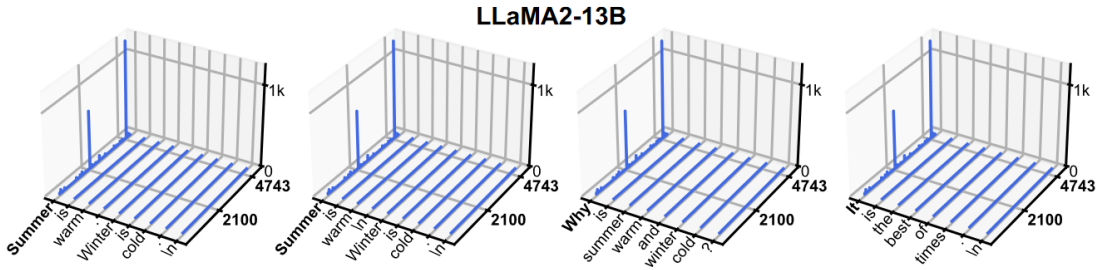


Figure 2: **Massive activations in LLaMA2-13B.** In this model, they appear in two fixed feature dimensions (2100, 4743), and are limited to the starting token.

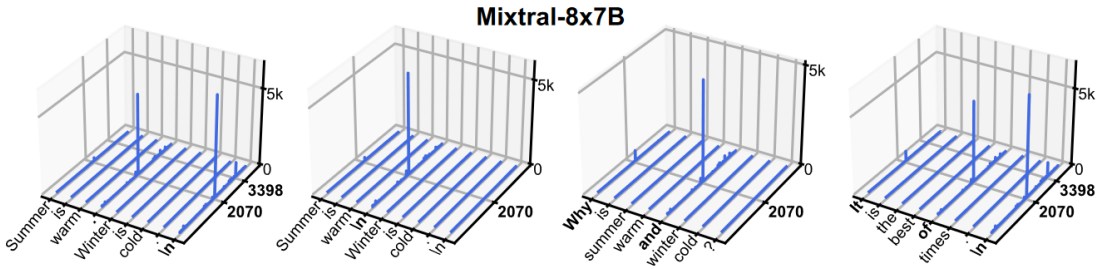


Figure 3: **Massive activations in Mixtral-8x7B.** In this model, they lie in two feature dimensions (2070, 3398), and are found within the starting token, delimiter tokens and certain word tokens ("`and`" and "`of`").

Massive Activations

- Properties:
  - rare & massive

Model	Top 1	Top 2	Top 3	Top 4	Top 5	Top-10	Top-100	Top 1%	Top 10%	median
LLaMA2-7B	**2622.0**	**1547.0**	**802.0**	**477.3**	156.9	45.7	10.6	1.1	0.6	0.2
LLaMA2-13B	**1264.0**	**781.0**	51.0	50.5	47.1	43.5	16.6	1.9	1.1	0.4
Mixtral-8x7B	**7100.0**	**5296.0**	**1014.5**	**467.8**	**302.8**	182.8	90.8	3.0	1.0	0.3

Table 1: Five largest, top 1% and 10%, and the median *activation magnitudes* at a hidden state of three LLMs. The activations that are considered as massive activations are highlighted in bold.

Massive Activations

- Properties:
  - rare & massive
  - fixed location
    - layer: Massive activations exist and remain as largely **constant** values throughout most of the **intermediate layers**. They **emerge** in the initial layers and start to diminish in the last few layers.

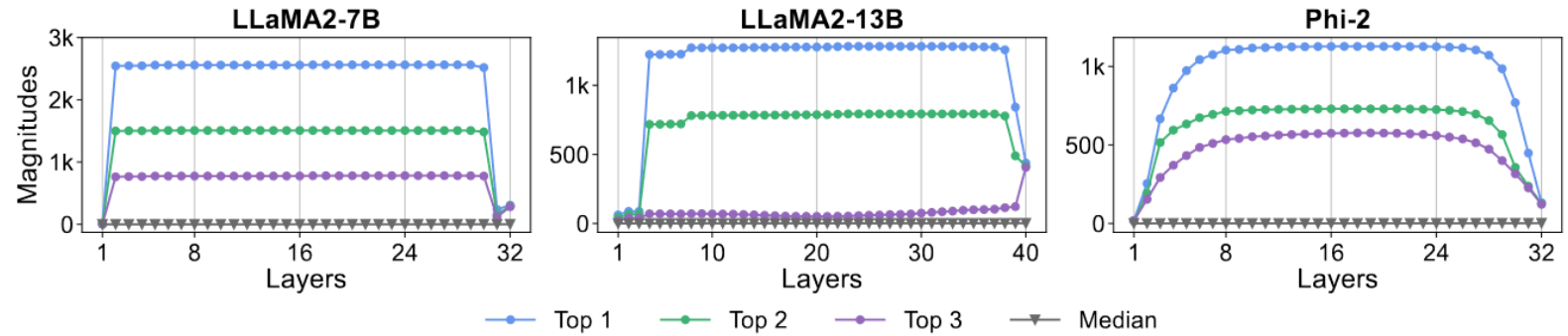


Figure 4: Three largest activation magnitudes and the median magnitude at each layer in LLMs.

Massive Activations

- Properties:
  - rare & massive
  - fixed location
    - layer: Massive activations exist and remain as largely **constant** values throughout most of the **intermediate layers**. They **emerge** in the initial layers and start to diminish in the last few layers.
    - channel: present in very few **fixed dimensions**
    - token:
      - a) Starting token only.
        - Models include LLaMA2-13B, MPT and GPT-2.
      - b) Starting token and the first “strong” delimiter token (i.e., “.” or “\n”)
        - Models include LLaMA2-7B and LLaMA2-7B-Chat.
      - c) Starting token, delimiter tokens (such as “.”, “\n”, “’” or “,”), and certain word tokens with weak semantics (such as “and”, “from”, “of” or “2”)
        - Models include LLaMA2-70B, Mistral-7B, Mixtral-8x7B, Falcon-40B and Phi-2.

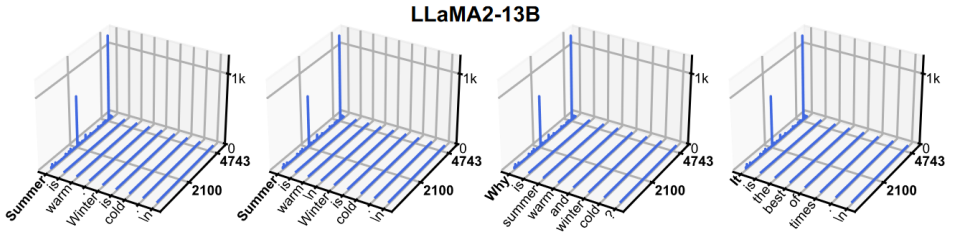


Figure 2: **Massive activations in LLaMA2-13B.** In this model, they appear in two fixed feature dimensions (2100, 4743), and are limited to the starting token.

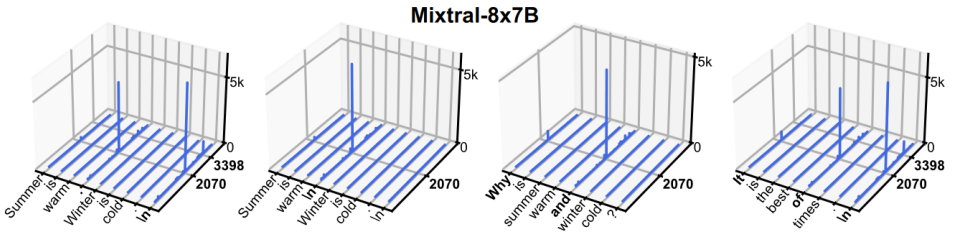


Figure 3: **Massive activations in Mixtral-8x7B.** In this model, they lie in two feature dimensions (2070, 3398), and are found within the starting token, delimiter tokens and certain word tokens (“`and`” and “`of`”).

Massive Activations

- Difference from Outlier Features [Dettmers et al. (2022)]
  - A massive activation is a scalar value, determined jointly by the sequence and feature dimensions; massive activations are present at extremely few token
  - An outlier feature is a vector, corresponding to activations at all tokens; outlier features expect most activations in them to be large
  - In LLaMA2-7B & 13B
    - a feature is deemed as an outlier feature if activation magnitudes exceed 6.0 at more than 25% of layers and 6% of tokens, on more than 90 out of 100 sequences
    - discover 10 and 25 outlier features in these two models respectively. However, none of them correspond to the feature dimensions of massive activations.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In NeurIPS, 2022.

Massive Activations Act as Biases in LLMs

- Are they important for internal computation? Or are they simply redundant activations with no effect?

Massive Activations Act as Biases in LLMs

- The variances of massive activations across input sequences

Model	Top 1	Top 2	Top 1%	Top 10%	Median
LLaMA2-7B	2556.8 ± 141.0	-1507.0 ± 83.0	-0.14 ± 0.6	0.0 ± 0.5	0.2 ± 0.3
LLaMA2-13B	-1277.5 ± 14.6	-787.8 ± 8.0	0.9 ± 0.7	-0.3 ± 0.8	-0.3 ± 0.6

Table 2: The mean and variance of activation values at several positions, corresponding to the 2 largest, top 1% and 10%, and the median magnitudes within the hidden state. We find that the variation in massive activations is significantly lower in comparison to other activations.

- the variances of massive activations are considerably smaller relative to their mean values when compared to other activations.

-> suspect that the accurate value of massive activations is not important

Massive Activations Act as Biases in LLMs

- Experiment: modify the inference of LLMs by intervening massive activations at one layer
  - Experimental setup
    - LLaMA-7B&13B
    - evaluation: ppl on WikiText, C4, PG-19; zero-shot accuracy on BoolQ, PIQA, WinoGrande, Arc-Easy & Arc-Challenge
    - perform interventaion once on the hidden state where massive activates first appear
  - Result

	LLaMA2-7B				LLaMA2-13B			
Intervention	WikiText	C4	PG-19	Mean Zero-Shot	WikiText	C4	PG-19	Mean Zero-Shot
Original	5.47	7.85	8.57	68.95%	4.88	7.22	7.16	71.94%
Set to zero	`inf`	`inf`	`inf`	36.75%	5729	5526	4759	37.50%
Set to mean	5.47	7.86	8.59	68.94%	4.88	7.22	7.16	71.92%

    - Massive activations act as fixed but important biases in LLMs

Effects on Attention

- Attention is Concentrated on Massive Activations
  - Figure 5 shows the attention logits (before softmax), averaged over all heads per layer in LLaMA2-7B
  - many attention logits tend to be negative following massive activations
    - mostly computed by the inner product between query and key states of tokens without massive activations
    - the key states belong to tokens associated with massive activations, the resulting attention logits are slightly positive.

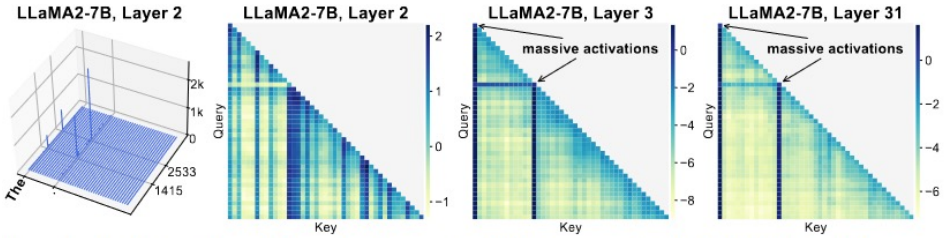
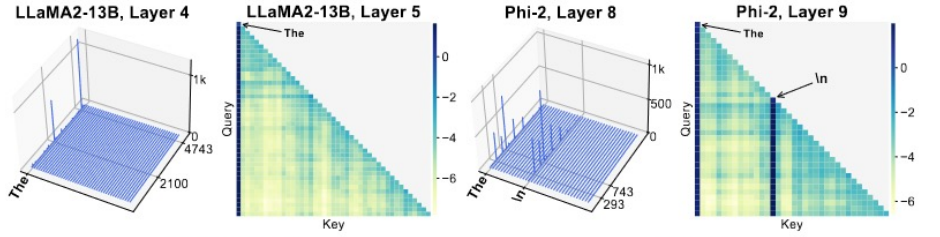


Figure 5: Attention patterns *before* and *after* massive activations appear in LLaMA2-7B. For each layer, we visualize average attention logits (unnormalized scores before softmax) over all heads, for an input sequence.

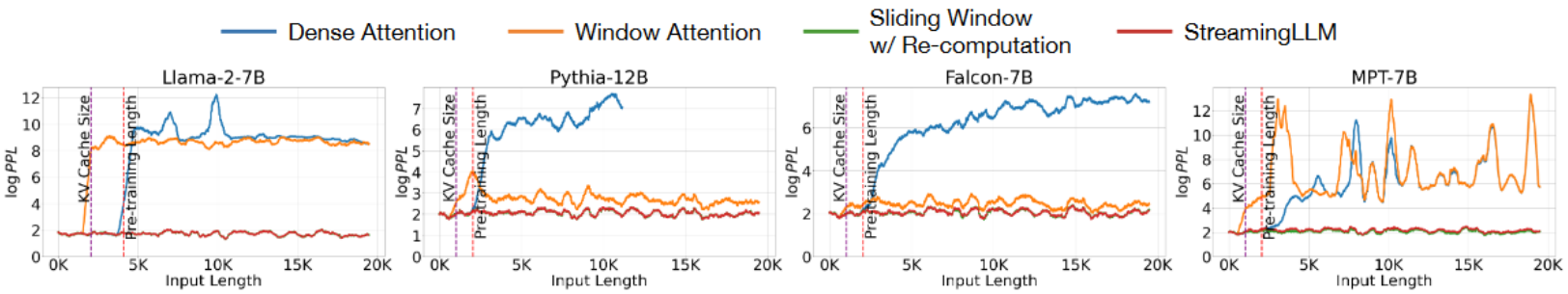


(a) LLaMA2-13B (b) Phi-2

Figure 6: Attention patterns *after* massive activations emerge in LLaMA2-13B (*left*) and Phi-2 (*right*).

Effects on Attention

- Attention Sink(Xiao et al., 2023b)
  - The model collapses once the sequence length exceeds the cache size, i.e., even just evicting the KV of the first token
  - attention sink: Alongside the current sliding window tokens, we reintroduce a few starting tokens' KV in the attention computation.



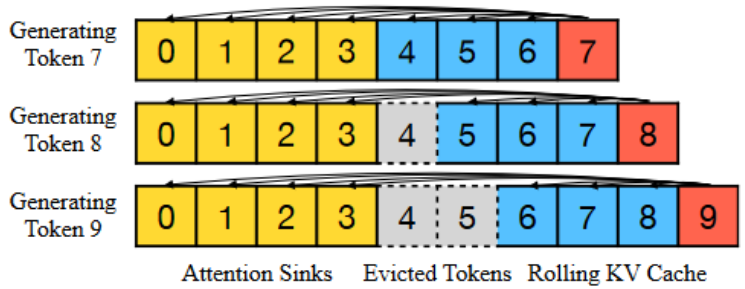
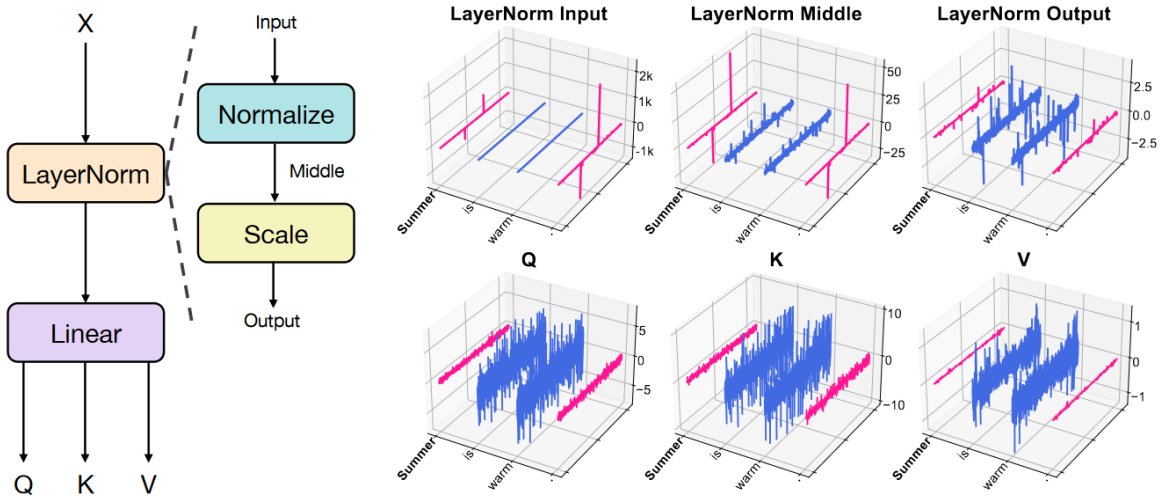


Figure 4: The KV cache of StreamingLLM.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. arXiv:2309.17453, 2023b.

Effects on Attention

- Massive Activations Impose Implicit Attention Biases
  - Attention LayerNorm & QKV projections
    - the subsequent QKV states exhibit considerably smaller variations within each embedding



(a) Attention LayerNorm and QKV linear projections.

(b) Layer 3, LLaMA2-7B. We highlight the embeddings of the two tokens where massive activations appear: the starting token and the period token.

Figure 7: Activation trajectory starting from input hidden states to query, key and value states.

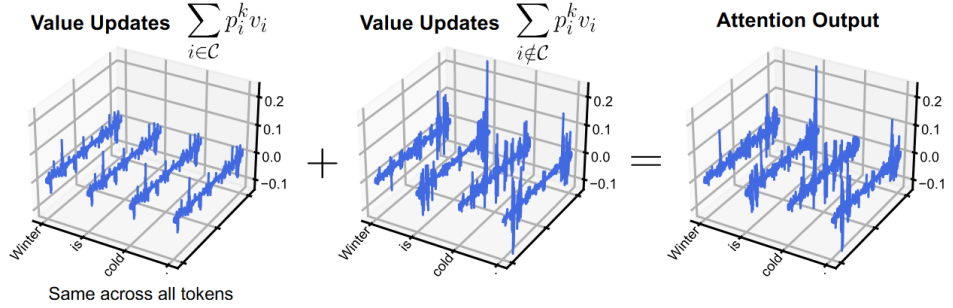
Effects on Attention

- Massive Activations Impose Implicit Attention Biases
  - Attention output decomposition
    - We decompose the attention output at each token ***k*** into two parts: value updates from the tokens ***C*** where attention is concentrated; and value updates aggregated from other tokens:

$$\text{Attention}(Q, K, V)_k = \sum_{i \le k} p_i^k v_i = \sum_{i \in \mathcal{C}} p_i^k v_i + \sum_{i \notin \mathcal{C}} p_i^k v_i$$

where p^k_i is the attention distribution of query token ***k*** to token ***i***

  - value updates from ***C*** are nearly identical across tokens



-> LLMs use massive activations to allocate substantial attention at certain tokens. These tokens are then utilized to form a constant bias term when computing the attention output.

Effects on Attention

- Explicit Attention Biases Eliminate Massive Activations
  - introduce bias terms to augment self-attention leads to no massive activations
    - Formulation: add additional learnable parameters k' v'

$$\text{Attention}(Q, K, V; \mathbf{k}', \mathbf{v}') = \text{softmax}\left(\frac{Q\left[K^T \;\; \mathbf{k}'\right]}{\sqrt{d}}\right)\begin{bmatrix} V \\ \mathbf{v}'^T \end{bmatrix}$$

    - Result

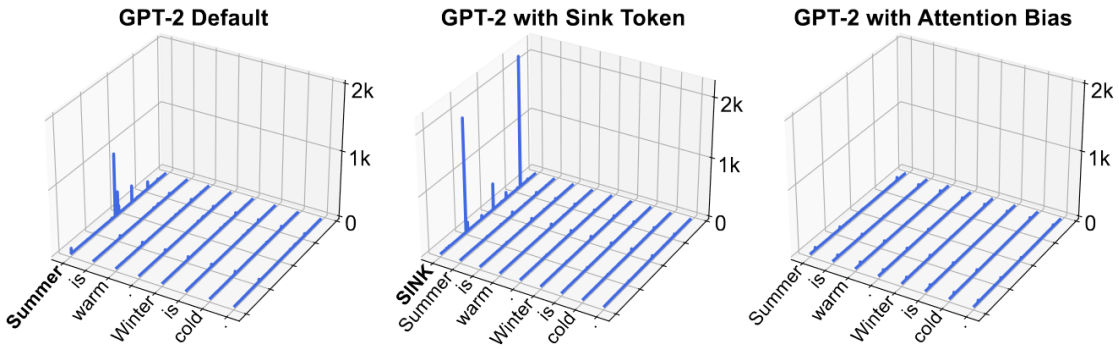


Figure 9: Massive activations disappear when training GPT-2 with explicit attention bias (Equation 3).

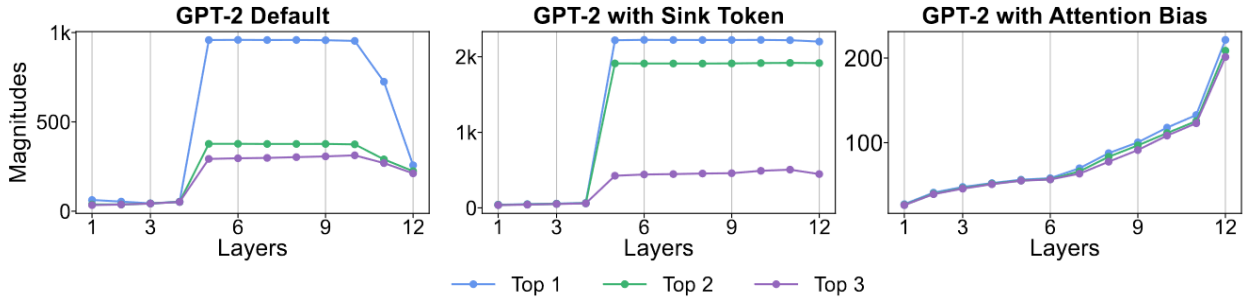


Figure 10: Three largest activation magnitudes in the output feature of each layer for three GPT-2 models.

Effects on Attention

> *Massive activations are connected to self-attention.*
>
> *LLMs use massive activations to concentrate substantial attention on very few tokens, injecting implicit bias terms in the attention computation.*
>
> *Further, massive activations can be eliminated by augmenting LLMs with explicit attention biases.*

Massive Activations in Vision Transformers

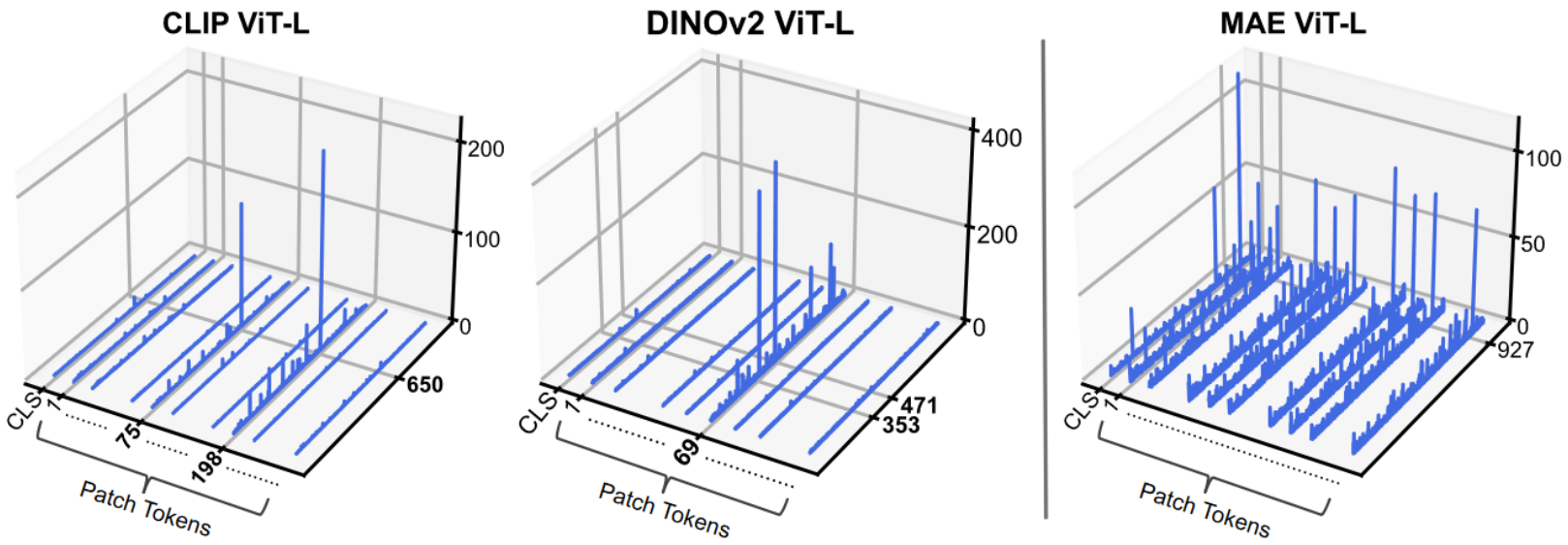


Figure 11: Massive activations are present in ViT-L from CLIP and DINOv2, but not MAE.

- Massive activations exist inCLIP ViT & DINOv2 VIT-L
- However, in MAE ViT-L, a feature dimension (927) exhibits uniformly large values across all tokens

Massive Activations in Vision Transformers

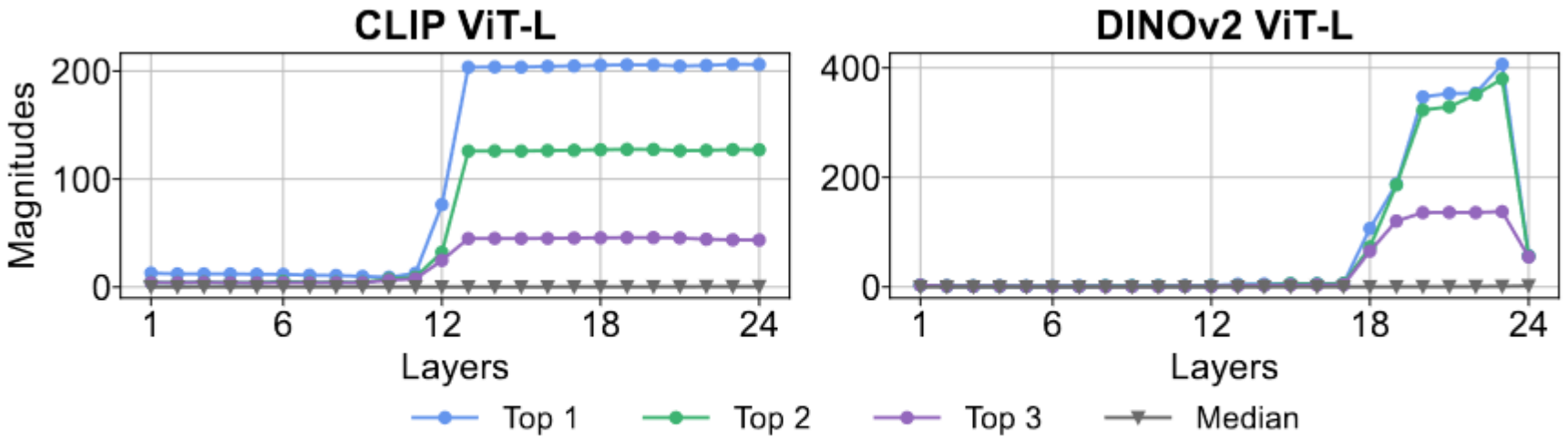


Figure 13: Three largest activation magnitudes and the median magnitude at each layer in CLIP and DINOv2 ViT-L.

- Massive activations are biases in ViTs
  - massive activations are consistently present across images and their values remain largely the same around the mean values.
  - unlike LLMs, massive activations start to appear only in the later stages of ViTs

Massive Activations in Vision Transformers

CLIP ViT-L, layer 13	
Intervention	ImageNet acc (%)
Original	75.5
Set to zero	59.8
Set to mean	75.5

Table 4: Intervention analysis of massive activations in CLIP ViT-L.

- Massive activations are biases in ViTs
  - setting massive activations to zero leads to significant drop in accuracy
  - while setting to their means results in negligible accuracy drop

Massive Activations in Vision Transformers

- Registers[Darcet et al. (2023)] are biases in ViTs
  - training ViTs with register tokens leads to smooth attention maps, and the resulting model family, namely DINOv2-reg, achieves superior downstream performance over DINOv2.
  - registers aggregate global input information

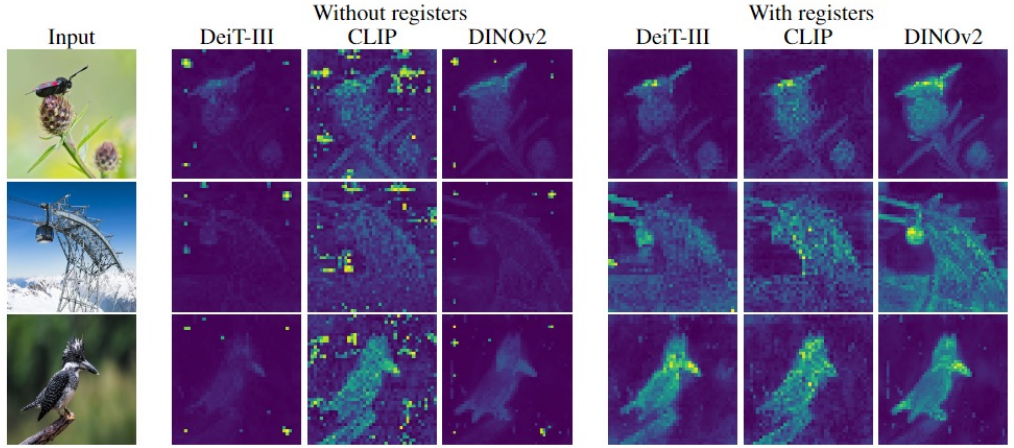
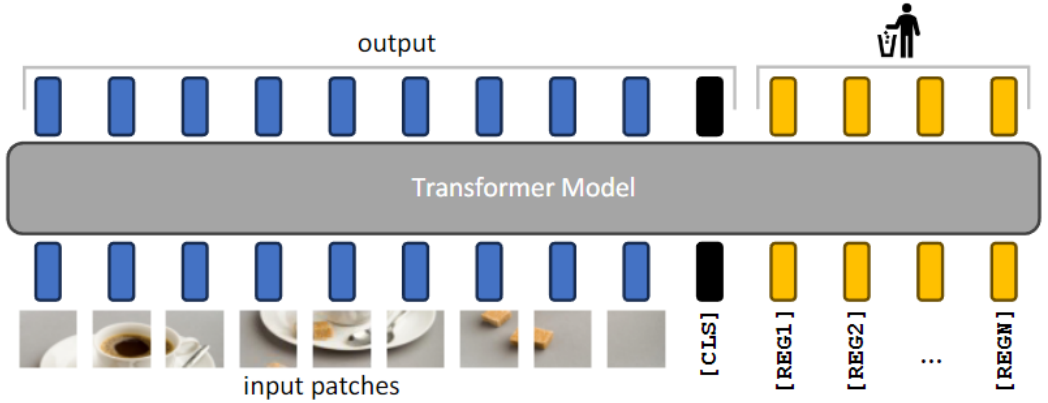


Figure 1: Register tokens enable interpretable attention maps in all vision transformers, similar to the original DINO method (Caron et al., 2021). Attention maps are calculated in high resolution for better visualisation. More qualitative results are available in appendix D.



Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. arXiv:2309.16588, 2023..

Massive Activations in Vision Transformers

- Registers[Darcet et al. (2023)] are biases in ViTs
  - Examining the largest ViT-G model in DINOv2-reg, we observe the existence of massive activations, as shown in Figure 14
  - Figure 16 visualizes the attention distribution of the [CLS] token in the last layer. We find that most of the attention is allocated to register 3
  - replace all register features at the output of every layer with their means, averaged over 10k ImageNet training images.

-> registers are not global input information aggregater

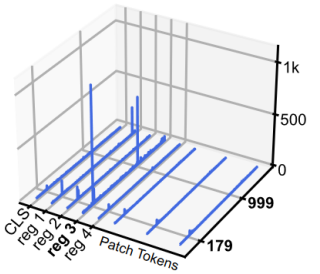


Figure 14: DINOv2-reg ViT-G.

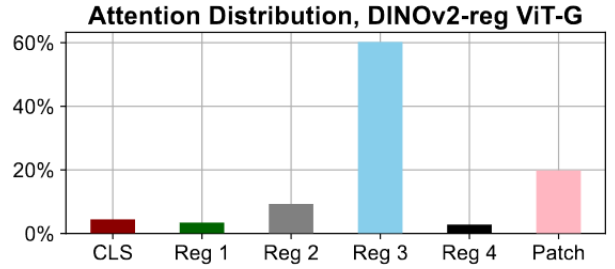


Figure 16: Average attention of the [CLS] token.

	DINOv2-reg with 4 registers			
ImageNet acc (%)	ViT-S	ViT-B	ViT-L	ViT-G
Original	81.9	84.8	86.3	87.0
`Fix-Reg-Mean`	81.7	85.0	86.2	87.0

Table 5: We fix *all* register features at *every* layer to their means and evaluate the intervened ViTs.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. arXiv:2309.16588, 2023..

Massive Activations in Vision Transformers

> *Massive activations exist in many but not all ViTs.*
>
> *Similar to those in LLMs, these activations act as constant biases.*
>
> *We also show the recently proposed register tokens have a similar function.*