# A 0-shot Self-Attention Mechanism for Accelerated Diagonal Attention

Viti Mario
Cyclope.ai
mario.viti@cyclope.ai

Nadiya Shvai
Cyclope.ai
nadiya.shvai@cyclope.ai

Arcadi Llanza
Unviersité Paris Est Créteil
arcadi.llanza-carmona@u-pec.fr

Amir Nakib
Unviersité Paris Est Créteil
nakib@u-pec.fr

## Abstract

*The ability of Transformers to process longer sequences has led to unprecedented levels of generalization in visual tasks. However, the complexity of Transformers is dominated by the quadratic cost associated with the computation of the attention blocks, posing a bottleneck that impedes the scaling of sequence length and the realization of more advanced AI solutions. We propose and explore the hypothesis that the self-attention mechanism exhibits regularities that can be exploited to enhance performance and achieve linear-cost attention without significant loss of effectiveness. Specifically, we investigate the attention matrix of Visual Transformers to identify and leverage these regularities in order to simplify the computation process. The resulting procedure significantly reduces the computational cost of Transformers by directly reducing attention block complexity. Moreover, the designed procedure is 0-shot self-supervised, thus it requires no retraining, additional data or parameters, as all Transformer parameters remain unchanged. Consequently, the proposed method can be seamlessly applied to pre-trained Visual Transformers without the need for retraining. Experiments conducted on a series of Vision Transformers pre-trainedon ImageNet-1K dataset demonstrate the effectiveness of our proposed approach.*

## 1. Introduction

The Visual Transformer (ViT) model has emerged as a novel architecture in the field of computer vision, demonstrating new state-of-the-art performance in a variety of tasks, from image classification to object detection and segmentation [5, 16]. These results can be attributed to the ability to capture long-range dependencies within images through self-attention mechanisms, enabling effective feature extraction and representation learning [17].

However, one of the key aspects of model performance lies in the need for longer sequences of data input to capture intricate patterns and dependencies within complex datasets. Longer sequences enable Transformers self-attention mechanism to exhibit enhanced understanding and reasoning capabilities, leading to improved performance across various tasks. Yet, this pursuit of longer sequences comes with a quadratic increase in computational complexity. As the length of input sequences grows, the number of pairwise interactions that need to be computed in the self-attention mechanism of Transformers escalates in quadratic way, resulting in a surge in computational requirements posing a significant bottleneck and potential limitations to more practical uses e.g. real-time processing [3, 9, 21].

In this paper, we explore various techniques and methodologies aimed at reducing the complexity of ViTs while preserving their efficacy across different tasks. We discuss the implications of these advances on both theoretical understanding and practical applications, shedding light on the potential impact of streamlined ViT models in addressing real-world challenges. Through our analysis and experimentation, our aim is to contribute to ongoing efforts to make ViT models more efficient, scalable, and applicable to a diverse range of domains and scenarios.

Our main contributions can be resumed as follows.

- A theoretical analysis on the diagonality of attention matrices.
- A method to approximate attention as almost diagonal matrix.
- A simple self-supervised procedure that can be applied to any Transformer.

The remainder of the paper is organized in this manner. Section 2 introduces the related work. Section 3 presents the methodology of the experiments conducted. Section 4 explains the experimental setup and presents the obtained results. Finally, Section 5 concludes the paper.

## 2. Related Works

Although there is still debate among researchers whether ViTs consistently outperform Convolutional Neural Networks [4, 5, 13], addressing the quadratic bottleneck in Transformer architectures has been a focal point of research in recent years, leading to the development of various approaches aimed at mitigating computational complexity while preserving model efficacy. These approaches can broadly be categorized into two main strategies: Approximation, and Global/Local Attention.

### 2.1. Approximation

This approach involves approximating self-attention matrices by exploiting their regularities and sparsity [18, 19]. Matrix approximation techniques such as low-rank decomposition, structured sparsity, hashing-based, and clustering-based methods have been explored to approximate attention matrices while maintaining the discriminative power of Transformers. By compressing attention matrices, these methods enable more efficient processing of longer sequences, making Transformers more scalable and applicable to a wider range of tasks. Another approach consists in reducing the cost of attention by approximating non-linear softmax activation by means of a linear attention mechanism [7]. By exploiting numerical similarities between softmax and Relu attention, and thus avoiding the computation of exponential functions, recently [2, 11] achieved SOTA in many ViT-based tasks.

### 2.2. Local and Sparse Attentions

Additionally, researchers have proposed strategies to limit the range of interactions in the self-attention mechanism by introducing local attention patterns. Instead of attending to all tokens in the input sequence, local attention mechanisms restrict attention to a subset of tokens within a certain neighborhood, thereby reducing the computational cost of attention computation. Various formulations of local attention, including window-based and kernel-based approaches, have been investigated to balance computational efficiency with modeling capacity [1, 8, 10, 15]. By incorporating local attention patterns, Transformers can effectively capture both short- and long-range dependencies in input sequences while mitigating the quadratic bottleneck associated with global attention. However, these methods apply specifically to visual tasks as they exploit the notion of spatial neighboring visual features being highly correlated within object representation boundaries; thus they do not transfer for all Transformer applications [12, 14].

## 3. Methodology

While the strategies discussed for mitigating the quadratic bottleneck in Transformers are effective, they often come with a significant caveat – the requirement for retraining or architectural adaptation. This necessity presents a practical hurdle, especially in scenarios where time, resources, or annotated data are scarce. To circumvent these challenges, we propose an innovative zero-shot methodology that capitalizes on the knowledge embedded within pretrained Vision Transformers (ViTs). Our approach diverges from conventional methods by exploiting the latent representations already encoded in ViTs, facilitating seamless integration with existing Transformer architectures without the need for retraining or architectural overhauls.

### 3.1. Transformer Attention Mechanism

The input sequence is initially embedded into vectors. Each token in the sequence is represented by a vector, where each dimension corresponds to a learned embedding feature. For each token representation in the input sequence, three vectors are derived: query vector, key vector, and value vector. These vectors are obtained through learned linear transformations of the input embedding. The attention mechanism computes scores between pairs of tokens in the sequence. The attention score between a query token $Q$ and a key token $K$ is computed using a dot product of their respective query and key vectors scaled by a factor $\sqrt{d_k}$ and normalized using the softmax function.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The attention weights are then used to compute a weighted sum of the value vectors associated with each key, this operation aggregates information from all tokens in the sequence, with each token's contribution weighted by its attention weight. In the Transformer architecture, as an optimization, attention is typically performed in parallel multiple times, each with its own set of learned query, key, and value transformations (multi-head attention).

### 3.2. Almost diagonal matrices

An almost diagonal matrix is a square matrix where most of the elements are concentrated along the main diagonal and a few adjacent diagonals. Diagonal matrices are close to their eigendecomposition and singular value decomposition [6]. To estimate the closest of a matrix diagonal values and eigenvalues one can use classical perturbation results for Hermitian matrices: given $A = A^*$ and $B = B^*$.

$$|\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|_2, \quad 1 \leq i \leq n \qquad (1)$$

Where $\lambda_i(X)$ are the eigenvalues of $X$. By choosing $B = diag(A)$, where $\text{diag}(X)$ is the diagonal matrix with diagonal entries of $X$, we obtain:

$$|\lambda_i(A) - a_{ii}| \leq \|A - \text{diag}(A)\|_2, \quad 1 \leq i \leq n \tag{2}$$

Thus, an upper bound of the closeness of a eigenvalue to the corresponding diagonal entry is given by the spectral norm of the matrix. It can be demonstrated [6] that for a given value $|A-\text{diag}(A)\|_2 = \epsilon$ the relative gap between the eigenvalues and its diagonal value $|\lambda_i(A) - a_{ii}| \,/\, |\lambda_i(A)|$ is of the order $\epsilon^2$, this holds if:

$$\|A - \text{diag}(A)\|_2 < \min_{i \neq j} |\lambda_i - \lambda_j| \tag{3}$$

The minimum gap as the right-hand side of (3) becomes the upper-bound of the diagonal estimation where $A$ is almost diagonal, thus gives us a method to decide whether a matrix can be considered almost diagonal. These properties hold for SVD as well with some slight deviation the homologous of (1) from Hermitian perturbation classical results from the singular values $\sigma$,

$$|\sigma_i(A) - |a_{ii}|| \leq \|A - \text{diag}(A)\|_2, \quad 1 \leq i \leq n \tag{4}$$

similarly for (3)

$$\|A - \text{diag}(A)\|_2 < \min_{i \neq j} |\sigma_i - \sigma_j| \tag{5}$$

#### 3.2.1 Almost diagonal attention

The self-attention mechanism in the Transformer architecture is a key component responsible for capturing contextual dependencies within a sequence of tokens. It enables the model to weigh the importance of each token in the sequence with respect to every other token, allowing for efficient processing of long-range dependencies.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Where $Q, K, V = W_q X, W_k X, W_v X$ are linear projections of the input sequence $X \in \mathcal{R}^{F_i \times N}, W_{q,k,v} \in \mathcal{R}^{F_o \times F_i}$

Inequality (4) holds for Hermitian symmetric matrices. At the base of the attention matrix there is the query key product and especially for self-attention,

$$QK^T = W_q X (W_k X)^T = W_q X X^T W_k^T \tag{7}$$

The product $QK^T$ is Hermitian if $W_k = W_q$ as follows from $XX^T = (XX^T)^T$. The lower bound (2) holds obviously for $QK^T$ as outer-product matrices share singular values and eigenvalues, the upper bound should be checked for any given value $X$ but as the matrix $QK^T$ is expressed as a function of $X$ unknown data, while $W_k, W_q$ are learned parameter, and these are fixed for a pre-trained model during learning. Thus an upper bound of gap between using the diagonal of the attention matrix instead of the whole attention can be computed as,

$$\begin{aligned}
\|QK^T - \text{diag}(QK^T)\|_2 = \\
\|QK^T_{i \neq j}\|_2 = \\
\|W_q (XX^T)_{i \neq j} W_k^T\|_2 = \\
\|W_q (XX^T - \text{diag}(XX^T)) W_k^T\|_2 \leq \\
\|W_q\|_2 \|W_k\|_2 \|XX^T - \text{diag}(XX^T)\|_2
\end{aligned}$$

Thus if $W_q \sim W_k$ the attention matrix is almost diagonal if

$$\begin{aligned}
\|QK^T - \text{diag}(QK^T)\|_2 < \\
\|W_q\|_2 \|W_k^T\|_2 \min_{i \neq j} \left|\sigma_i(XX^T) - \sigma_j(XX^T)\right|
\end{aligned}$$

We know that if (5) hold than the matrix is considered almost diagonal, thus we can formulate a probability function of a matrix being almost diagonal

$$\begin{aligned}
P(QK^T \sim \text{diag}(QK^T)) = P(\|QK^T - \text{diag}(QK^T)\|_2 < \min_{i \neq j} |\sigma_i - \sigma_j|) = \\
P(\|W_q\|_2 \|W_k\|_2 \|XX^T - \text{diag}(XX^T)\|_2 < \min_{i \neq j} |\sigma_i - \sigma_j|)
\end{aligned} \tag{8}$$

Because we have no control on the data $X$ but we have full control on the parameters $W_k, W_q$ we need to satisfy 2 criteria conjointly.

- choose $W_q, W_k$ that minimize $\|W_q\|_2 \|W_k\|_2$ to increase the probability when $X$ is unknown as the error upper bound in (eq. 8) is proportional to this value.
- choose $W_q, W_k$ so that $W_k \sim W_q$ as symmetry is a necessary condition of diagonality.

We will focus on minimizing the first term of the upper bound by checking whether $W_k \sim W_q$, we maximize the known term and obtain an almost diagonal attention, as an experiment we observe on data whether the method has em-

pirical validation by explicitly computing the upper bound with data. In order to make calculation feasible we will focus on a per head test to obtain an almost diagonal attention. To assess the viability of employing diagonal attention within individual attention heads of Transformer blocks, we devise a straightforward test aimed at determining whether a head can be accurately approximated using solely its diagonal values. The test involves computing the diagonal approximation error and comparing it against an upper bound (5) suitable for nearly diagonal matrices, as it turn out this involves checking whether $W_k \sim W_q$. By establishing an upper threshold $\epsilon$, we can reliably determine whether a head qualifies for diagonal attention. This 0-shot approach enables the provision of a set of indices per block that denote which heads are eligible for computation using diagonal attention.

### 3.3. Static diagonality test

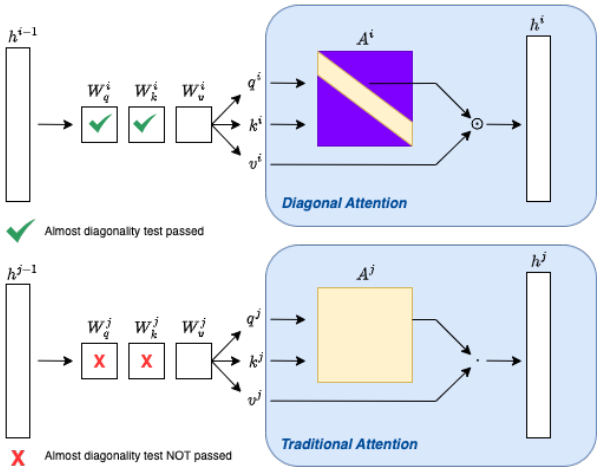



Figure 1. Diagram visualization of diagonal attention. The static diagonality test is executed once before inference for each attention block's parameters $W_k, W_q$ (where $i, j$ are block indices).

By static we mean that no extra data is used for this method, only the pre-trained parameters thus 0-shot. The scaling factor of the error on the attention matrix is $\|W_k\|_2\|W_q\|_2$ by means of the analysis of diagonal values of $W_kW_q$ we aim both to minimize this factor and at the same time satisfy the constraint $W_k \sim W_q$.

The test will induce a binary choice, a head attention matrix will be applied diagonally, thus reducing drastically the computational complexity, for a head not passing the *diagonality* test the attention will be computed using the matrix vector product:

- Test on $W_qW_k^{T^i}$ passed : $h^i = \text{diag}(A^i) \odot v^i$, $i \in [0 - \#\text{Heads}[$
- Test on $W_qW_k^{T^j}$ not passed : $h^j = A^j v^j$, $j \in [0 - \#\text{Heads}[$

The test consist in evaluating the following constraint $\|W_k\|\|W_q\|\|W_k - W_q\| \leq \alpha\epsilon$, where $\epsilon$ is computed from as $\epsilon = \max_{l\in[0-L]} \|W_k^l\|\|W_q^l\|\|W_k^l - W_q^l\|$. Here $l$ is the depth index of the layer, $L$ is the index of the latest attention block as shown in diagram 1 and $\alpha \in [0 - 1]$ tunes the tradeoff between performance and accuracy. We observed that a higher $\alpha$ will "diagonalize" the attention of deeper and deeper layers. This observation correlates with the assumption that deeper layers of the Transformer model become increasingly "entangled", as the discrepancy between the matrix product and its diagonal form tends to grow with the depth of the Transformer. (Figures 2,3).

We will also empirically validate this static analysis of the parameters by observing the corresponding attention matrix showing similar *diagonality* trends (Figures 4,5).

## 4. Experiments and results

### 4.1. ImageNet-1K

In order to evaluate the effectiveness of our proposed method for selecting diagonal attention compared to existing approaches, such as the Swin Transformer V2 (both B and T versions), we will conduct experiments on the ImageNet-1K dataset using pre-trainedmodels.

- $B_{16}$ : the (BViT) base visual image Transformer [5] will be our baseline.
- $B_{16}$ \(Ours): adding diagonal attention (proposed method) to the baseline.
- $B_{16}$ 0.875 : shrinking minimally (by a factor of 0.875) the input size to obtain a slight improvement on the speed.
- $B_{16}$ 0.875 \(Ours): adding diagonal attention (proposed method) to the shrunk input baseline.
- Swin B v2 : an improvement on (ViT) proposed by [12] using local attention (B for base).
- Swin T v2 : an improvement on (ViT) proposed by [12] using local attention (T for tiny).
- kMaX-DeepLab [20] : a method to obtain cross attention using k-means clustered feature vectors, the attention matrix dimension is thus reduced based on the chosen k .
- DSA90% [10] : Dynamic Sparse Attention method proposes to accelerate inference by focusing on attention values, thus only the most relevant part of the attention is computed by creating a dynamic computational path, 90% indicates that 10% of the attention is used.

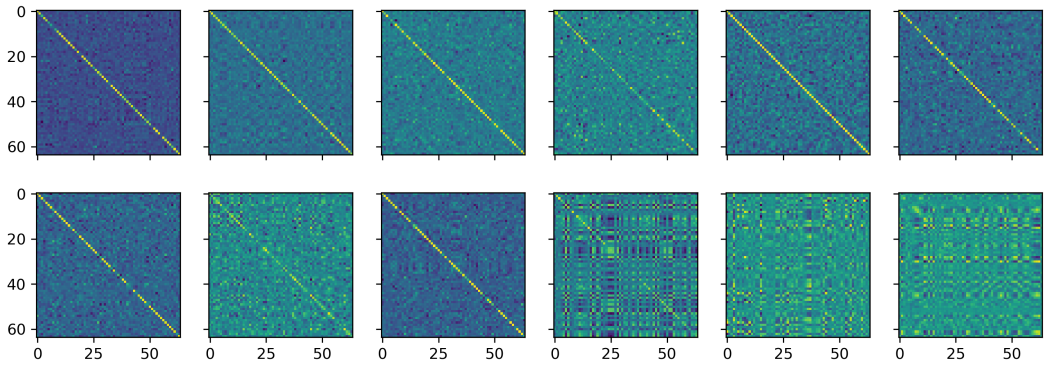


Figure 2. Visualizations of the attention matrix computed on the data sequence indicate that certain heads within the first layer of the Transformer model show correlated query-key features (values in $\mathbb{R}$). This correlation is apparent through elevated values along the diagonal elements of these matrices. Such observations suggest that specific heads in the Transformer model are inclined to focus on capturing and attending to related features within the input sequence, potentially reflecting a structured representation.

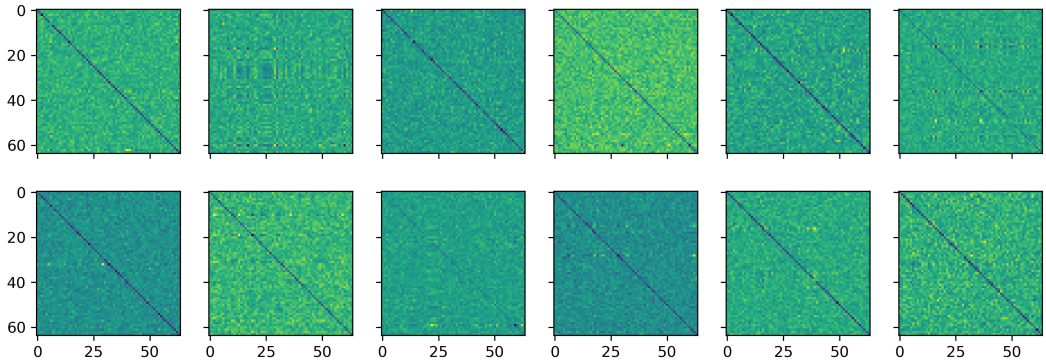


Figure 3. Examination of attention matrices per head within the final layer of the Transformer model reveals a notable phenomenon marked by an opposite correlation or *antidiagonality*. Specifically, analysis of the attention matrix computed on the data sequence demonstrates a pattern where query-key features display an inverse correlation, as evidenced by the predominance of lower values along the diagonal elements. This observation implies a distinct alteration in the information processing mechanisms of the model's higher-level attention layers compared to their earlier counterparts. Further investigation into these dynamics may yield valuable insights into the hierarchical information processing strategies adopted by Transformers.

Specifically, we will compare the performance of our method in terms of speed-up and performance degradation on both CPU and GPU platforms, considering the time and resource occupation per image (Fig.6 and Tab. 2).

Overall, our evaluation will provide insights into the trade-offs between performance and efficiency when employing our proposed method for selecting diagonal attention in Transformer models compared to state-of-the-art approaches like the Swin Transformer V2.

The results demonstrate that our method achieves significant speed-ups compared to state-of-the-art local attention methods, particularly the Swin Transformer, which is optimized for visual tasks. We also compared the top-1 accuracy of our method with that of Swin-B V2, as this provides the most equitable comparison Tab. 2. A key advantage of our method is that it operates in a zero-shot manner, meaning that it does not require fine-tuning or any modifications to the original architecture, unlike all other proposed methods. Although DSA achieves the most relevant speed-up and performance trade-off, such a method relies on custom operators for GPU acceleration, our approach does not necessitate these additional components, further simplifying its application.

### 4.2. Static diagonality test

The experiment produced significant findings, indicating a robust correlation between the diagonality of the correlation matrix of key-query parameters and the actual diagonality of attention matrices in Transformer models, static and dynamic diagonality test, respectively. This correlation remained consistent across various layers and attention heads within the Transformer model as shown for both

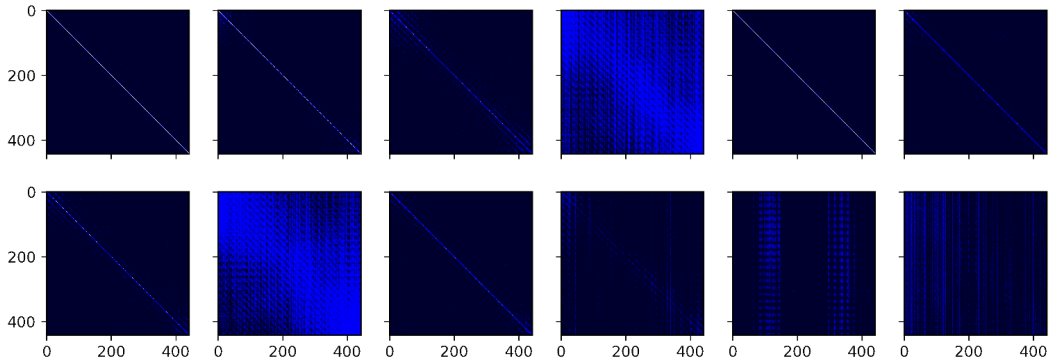


Figure 4. Visualizations of the attention matrices per head within the first layer of the Transformer model reveal notable patterns wherein certain heads exhibit correlated query-key features (values are normalized by softmax lies in the $(0-1)$ interval ). This correlation is evidenced by the heightened values along the diagonal elements of these matrices. Note: high contrast has been applied to address visualization issues.

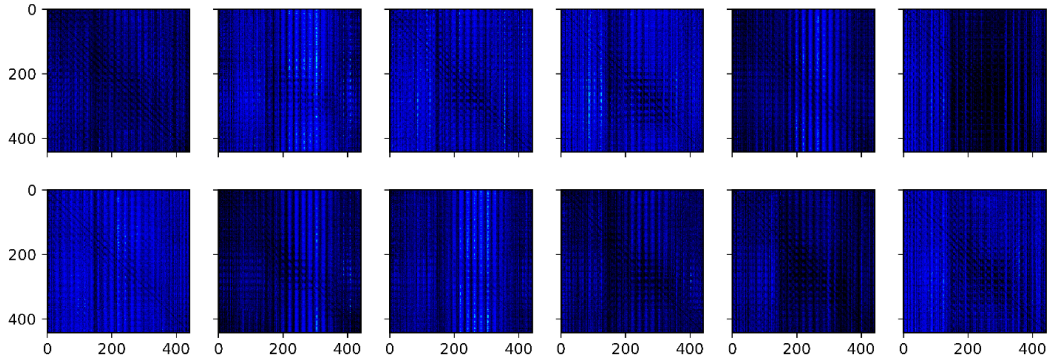


Figure 5. In contrast, analysis of attention matrices per head within the final layer of the Transformer model reveals a distinct phenomenon characterized by an opposite correlation or *antidiagonality*. Specifically, these visualizations showcase a pattern where query-key features are inversely correlated, evident from the prominence of lower values along the diagonal elements. This observation suggests a divergent behavior in the information processing mechanisms of the model's higher-level attention layers compared to their earlier counterparts. Further exploration of these dynamics may offer insights into the hierarchical information processing strategies employed by Transformers.

tested $W_q W_k^T$ (Fig. 2,3) and corresponding sample attention matrix (Fig. 4,5). Furthermore, our assessment of the upper bound error condition confirmed its efficacy in accurately identifying diagonal attention matrices. The condition reliably differentiated attention matrices with substantial levels of diagonality, aligning closely with the ground-truth evaluations Tab. 1. As an example, Figure (7b) shows the rule applied to the first layer of a pre-trained $B_{16}$ ViT model for the first attention block using the values of the last attention block. As a result, the last block will not be used, which is not a problem, as it will hold the most displaced information. In conclusion, these results validate the proposed method for selecting diagonal attention in Transformer models and underscore the importance of considering the relationship between key-query parameter correlations and attention matrix diagonality in Transformer architecture design.

Table 1. Almost diagonal analysis of the actual attention matrix averaged across a batch of images as shown in the method section. The rows correspond to the attention matrices $Blk_{\text{block id}} H_{\text{head id}}$. When spectral norm (3rd row) is below the minimal gap amongst all singular values(4rd row) the matrix can be considered almost diagonal [6]. These values refer to those shown in Fig. 7b for the choice of the diagonal attention head.

| Matrix | $\max_i \lvert\sigma_i - \lvert a_{ii}\rvert\rvert$ | $\lVert A - \operatorname{diag} A\rVert_2$ | $\min_{i\neq j} \lvert\sigma_i - \sigma_j\rvert$ |
|---|---|---|---|
| $Blk_0 H_0$ | 0.001 | 0.033 | 0.039 |
| $Blk_0 H_9$ | 0.013 | 0.134 | 0.031 |
| $Blk_{-1} H_0$ | 0.213 | 0.354 | 0.013 |

## 5. Discussion and conclusion

We presented a method for accelerating attention in Transformers, using the $B_{16}$ ViT architecture for comparison against state-of-the-art sparse attention methods, in-

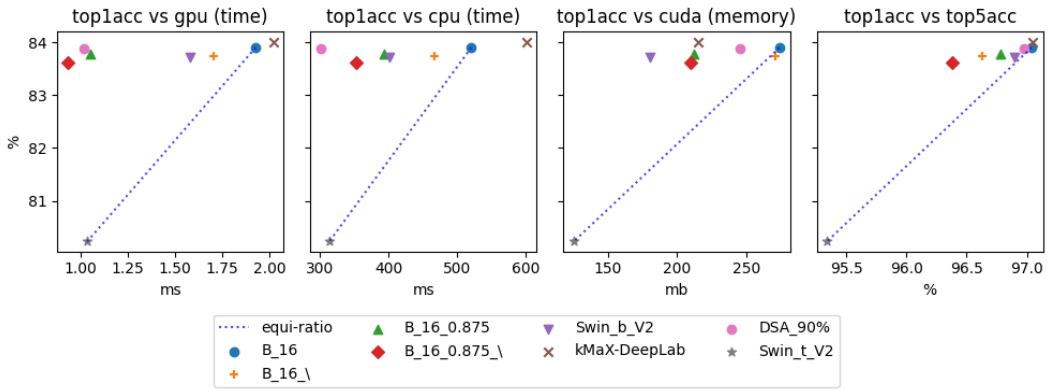
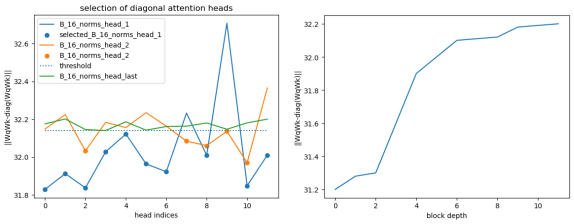



Figure 6. Metrics of Table 2 visualized in a scatter plot. The equi-ratio segment identifies all methods having the same compromise ratio for performances and accuracy, each point above this line represent a better compromise. The kMaX-DeepLab [20] althought being an attention decomposition method based on clustering does not provide any speed-up due to the k-means routine.

Table 2. Comparing our method to other sparse and local attention methods in literature. $B_{16}$ ViT pre-trained model is our baseline, $B_{16}$ \diagonal attention is applied, $B_{16}0.875$ the input is shrunk by a $0.875$ factor, $B_{16}0.875$ \diagonal attention is applied.

| Model | gpu ms | cpu ms | cuda mb | input size | top1acc% | top5acc% |
|---|---|---|---|---|---|---|
| $B_{16}$ [5] | 1.928 | 520.586 | 273.74 | (384, 384) | 83.90 | 97.04 |
| $B_{16}$ \(ours) | 1.708 | 465.051 | 270.37 | (384, 384) | 83.73 | 96.63 |
| $B_{16}$ 0.875 | 1.052 | 393.032 | 212.3 | (326, 326) | 83.78 | 96.78 |
| $B_{16}$ 0.875 \(ours) | **0.932** | 353.001 | 209.7 | (326, 326) | **83.62** | 96.38 |
| Swin b v2 [12] | 1.582 | 401.312 | 180.32 | (326, 326) | **83.71** | 96.90 |
| Swin t v2 [12] | 1.037 | 313.842 | 125.54 | (326, 326) | 80.23 | 95.34 |
| kMaX-DeepLab [20] | 2.021 | 600.831 | 215.32 | (326, 326) | 84.12 | 97.81 |
| DSA 90% [10] | 1.021 | 301.145 | 245.52 | (326, 326) | 83.88 | 96.99 |




(a) The static *diagonality* test as described in the method section for the $B_{16}$ model, in particular we observe the variability for the first, second and last attention block used as self-learned threshold.

(b) The *diagonality* measure here is shown averaged per head, it exhibit a growing pattern as for the hypothesis formulated that deep layer self-attention is more entangled to capture long term relationships.

Figure 7. Static *diagonality* test example on pre-trainedparameters on $B_16$ Visual Image Transformer.

cluding kMax-DeepLab, DSA, and Swin. Among these, the fairest comparison is with the Swin Transformer, as our focus is on image classification, and Swin has been specifically optimized for such visual tasks using windowed attention. On the other hand, DSA and kMax-DeepLab are more general-purpose attention optimizations.

Our method achieved comparable performance to these sparse attention approaches, while demonstrating a significant speed-up in both CPU and GPU inference, all at minimal cost. It is important to highlight the value of our method being zero-shot, as it requires neither retraining nor specific modifications to the original architecture. We also introduced a diagonality test, based on the theory of "almost diagonality," which has been empirically validated, as shown in Figures 2,3,4,5.

It is, however, essential to compare both the implementation and theoretical complexities. DAS (Dynamic Sparse Attention) benefits from a custom operator that leverages hardware optimizations, explaining its faster GPU inference times relative to our method. Furthermore, the number of operations required for the standard attention product,$B \times H \times N^2$, is significantly higher than for the diagonal attention product, which requires only $B \times H \times N$ operations. Yet, in practice, the time complexity gain is closer to a factor of 2 rather than $N$. This is due to the multi-head attention

mechanism in Transformers, which already mitigates some of the computational overhead.

Additionally, operations such as reshaping, while theoretically considered cost-free, can contribute significantly to the overall complexity in real-world implementations. Despite optimizations using libraries like einops, these operations still introduce computational overhead. Therefore, while theoretical optimizations offer insight, they may have limited impact in practical applications. It is crucial to account for both theoretical complexities and real-world implementation details when evaluating the efficiency of Transformers in applied settings.

## References


[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 2

[2] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Multi-scale linear attention for high-resolution dense prediction. *arXiv preprint arXiv:2205.14756*, 2022. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 4, 7

[6] V. Hari. Structure of almost diagonal matrices. *Mathematical Communications*, 4(2):135–158, Dec. 1999. 2, 3, 6

[7] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020. 2

[8] Jinpeng Li, Yichao Yan, Shengcai Liao, Xiaokang Yang, and Ling Shao. Local-to-global self-attention in vision transformers. *arXiv preprint arXiv:2107.04735*, 2021. 2

[9] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022. 1

[10] Liu Liu, Zheng Qu, Zhaodong Chen, Yufei Ding, and Yuan Xie. Transformer acceleration with dynamic sparse attention. *arXiv preprint arXiv:2110.11299*, 2021. 2, 4, 7

[11] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14430, 2023. 2

[12] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 2, 4, 7

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[14] Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. Do transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5758–5773, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[15] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018. 2

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1

[17] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[18] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *Advances in Neural Information Processing Systems*, 33:21665–21674, 2020. 2

[19] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2

[20] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. kmax-deeplab: k-means mask transformer. *arXiv preprint arXiv:2207.04044*, 2022. 4, 7

[21] Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C Knoll. Vision language models in autonomous driving and intelligent transportation systems. *arXiv preprint arXiv:2310.14414*, 2023. 1