Exploiting Information Redundancy in Attention Maps for Extreme Quantization of Vision Transformers

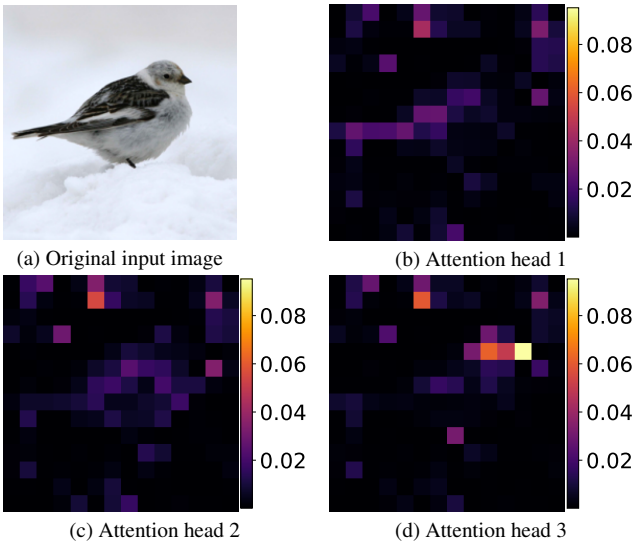
Lucas Maisonnave*
Université Paris-Saclay CEA, List
F-91120 Palaiseau, France
lucas.maisonnave@cea.fr

Karim Haroun*
i3S / CNRS, Université Côte d'Azur
Sophia Antipolis, France
karim.haroun@etu.univ-cotedazur.fr

Tom Pégeot
Université Paris-Saclay CEA, List
F-91120 Palaiseau, France
tom.pegeot@gmail.com

Abstract

Transformer models rely on Multi-Head Self-Attention (MHSA) mechanisms, where each attention head contributes to the final representation. However, their computational complexity and high memory demands due to MHSA hinders their deployment at the edge. In this work, we analyze and exploit information redundancy in attention maps to accelerate model inference. By quantifying the information captured by each attention head using Shannon entropy, our analysis reveals that attention heads with lower entropy, i.e., exhibiting more deterministic behavior, tend to contribute less information, motivating targeted compression strategies. Relying on these insights, we propose Entropy Attention Maps (EAM), a model that freezes the weights of low-entropy attention maps and quantizes these values to low precision to avoid redundant recomputation. Empirical validation on ImageNet-1k shows that EAM achieves similar or higher accuracy at $\leq$20% sparsity in attention maps and competitive performance beyond this level for the DeiT and Swin Transformer models.



(a) Original input image (b) Attention head 1 (c) Attention head 2 (d) Attention head 3

Figure 1. Visualization of the input image and the corresponding attention maps from the last layer's attention heads of DeiT-Tiny.

1. Introduction

Transformer models have achieved notable success in natural language processing (NLP) [34] and computer vision tasks [3, 6, 21, 30, 40], due to their ability to model long-range dependencies and handle variable-sized input sequences. Architectures such as BERT [5] and Vision Transformers (ViT) [6] have demonstrated state-of-the-art performance on a wide range of benchmarks. Their effectiveness is due to the Multi-Head Self-Attention (MHSA) mechanism, which allows the model to capture diverse contextual relationships through multiple attention heads operating in parallel.

However, Transformers impose significant computational and memory demands, essentially due to MHSA [32]. Computing attention requires pairwise interaction of all token embeddings, resulting in a quadratic complexity $O(N^2)$ with respect to the sequence length, where N is the number of tokens. This hinders deployment in environments with limited resources or tight latency requirements.

To mitigate these limitations, recent research has at-

*These authors contributed equally to this work.

tempted to reduce complexity by designing more efficient models. Amongst the developed techniques, two main categories of complexity reduction stand out. The first involves token-level sparsity [10], in which attention is selectively applied only to a subset of tokens. The second is quantization [15, 16, 18, 23, 37, 38], where model weights and activations are encoded with lower numerical precision (e.g., from FP32 to 4-bit integer), reducing both memory footprint and computational complexity.

In this work, we focus on precision reduction of attention heads within the MHSA module, relying on two main observations. First, our visual analysis of individual attention maps, shown in Figure 1 reveals that their weights frequently focus on small, localized regions of the input space rather than uniformly distributed across all positions. This spatial concentration suggests that a significant part of the attention computations may be redundant, as many attention weights contribute minimally to the context of the image. Second, we hypothesize that this redundancy can be quantified via Shannon entropy, i.e., heads exhibiting lower entropy, indicating limited variation across inputs, may be frozen and quantized to extremely low bit widths (as low as 4 bits) without affecting model performance, since their weights remain stable during inference. To test this hypothesis, we estimate for each head in every layer the entropy over the training dataset of the weights of the attention map. The resulting entropy values, as shown in Figure 2, allow us to identify heads with consistently low variability.

Relying on this observation, we develop a compression strategy that partially fixes (freezes) the attention weights of low-entropy attention maps and applies low-precision quantization. Despite removing their dynamic computation during inference, we retain the representational diversity of the high-entropy attention heads. Our contributions are summarized as follows:

- We apply an entropy-based measure to quantify the information of each attention head, revealing that all attention heads exhibit a variable entropy across inputs.
- We propose a model that partially fixes the attention weights of low-entropy attention maps during inference and applies 4-bit quantization, reducing the computational complexity and memory demands without altering the model performance.
- We conduct extensive experiments on ImageNet-1K across various ViT architectures against state-of-the-art, and validate these results with an ablation study.

2. Related Works

2.1. Vision Transformers

As discussed in the Introduction, ViTs introduced by Dosovitskiy et al. [6] rely on the self-attention mechanism that computes contextual relationships between all tokens, where each token's representation is generated through learned query, key, and value projections followed by softmax-weighted aggregation across the entire token sequence. Building upon ViT, subsequent work addressed efficiency and scalability limitations. DeiT [33] introduced distillation strategies to reduce training resource requirements, while Swin Transformer [22] proposed hierarchical feature maps and local-window attention to lower computational complexity. Despite these optimizations, the quadratic complexity of self-attention relative to the number of tokens in the sequence, coupled with high parameter counts, sustained significant computational and memory demands.

To mitigate these constraints, researchers have designed efficient architectures that target low computation and memory during inference. Among these, Swin Transformer [22] introduced a hierarchical design using shifted window partitioning to efficiently limit attention computation to local regions. Pyramid Vision Transformer (PVT) [36] adopted a progressive shrinking pyramid structure to handle high-resolution inputs with reduced computation.

Besides, specific model compression strategies have been developed for integration into existing architectures. These are broadly categorized into two families: token reduction, which exploits sparsity, and quantization, which reduces numerical precision. Token reduction methods decrease the sequence length processed by the ViT, either by pruning redundant tokens [19, 28, 31] or merging semantically similar tokens [2, 8, 9, 13]. These approaches reduce computational complexity, but leave model parameters uncompressed, resulting in comparable memory demands for weights. In this paper, we focus on quantization techniques that reduce the precision of both weights and activations. These will be detailed in the following section.

2.2. Quantization

Quantization reduces the numerical precision of weights and activations in neural networks, typically from 32-bit floating-point to lower bit-width fixed-point or integer representations [7]. Early quantization methods were developed for CNNs [11, 12], these methods include DoReFa-Net [41] which approximates gradients in quantization-aware training by straight-through estimator (STE) [1], and PACT [4] which propose parameterized clipping for activation quantization. Other works adopted non-uniform quantization [14, 24, 39], and mixed-precision quantization [26, 29, 35, 37], where different bit widths are assigned to weights and activations based on their sensitivity, typically determined through pre-computed measures prior to inference.

Recent research has extended quantization methods to Vision Transformers. Ranking Loss [23] preserves the relative order of quantized attention maps through a dedicated

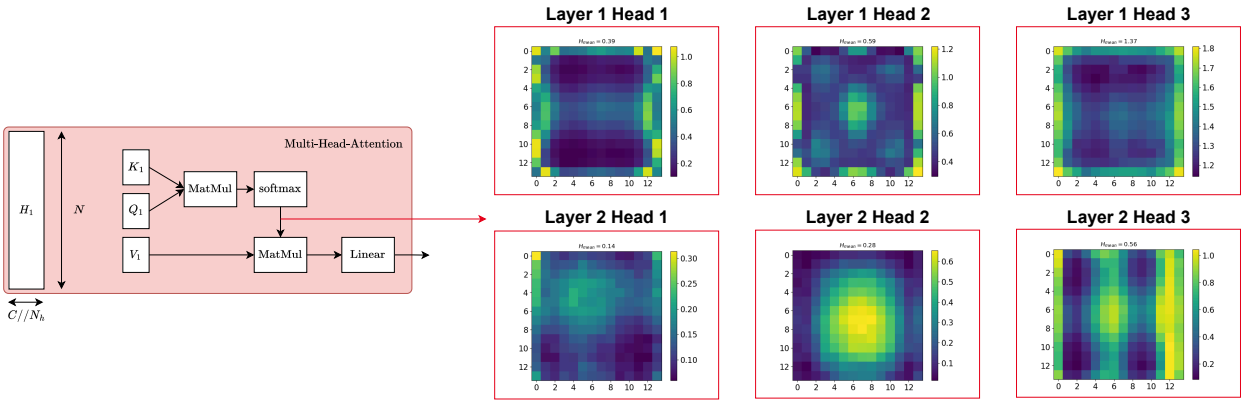


Figure 2. Entropy Attention Maps (EAM) of the CLS token of a DeiT-Tiny computed on 5% of ImageNet-1K. With this approach we can visualize differences between heads and the way they actually use information through attention.

loss function. Q-ViT [17] implements differentiable quantization, treating bit widths and scaling factors as learnable parameters during optimization. PTQ4ViT [38] introduces twin uniform quantization coupled with a Hessian-guided metric for scaling factor selection. To handle non-linear operations, FQ-ViT [20] employs powers-of-two scaling for LayerNorm and logarithmic integer quantization for Softmax outputs. RepQ-ViT [18] decouples quantization from inference pipelines to manage extreme activation distributions in LayerNorm and Softmax layers. Finally, PSAQ-ViT [16] enables data-free quantization by leveraging patch similarity metrics.

3. Motivations

Consider a tokenized input sequence $\mathbf{X} \in \mathbb{R}^{N \times d_e}$, with N tokens each of dimension d_e, obtained by partitioning an RGB image $I \in \mathbb{R}^{H \times W}$ into non-overlapping patches $\mathbf{I_p} \in \mathbb{R}^{H_p \times W_p}$. The Transformer architecture processes this through two modules, a Multi-Head Self-Attention (MHSA) and a Multi-Layer Perceptron (MLP). The MHSA mechanism first projects the input sequence X into three distinct representations query (Q), key (K) and values (V) through linear projections:

$$Q = \mathbf{X}W_Q, \quad K = \mathbf{X}W_K, \quad V = \mathbf{X}W_V \tag{1}$$

where $Q, K, V \in \mathbb{R}^{N \times d_e}$, and $W_Q, W_K, W_V \in \mathbb{R}^{d_e \times d_e}$ are learned projection matrices for queries, keys, and values, respectively. Attention is then computed using the scaled dot-product formulation:

$$A = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_e}}\right) \in \mathbb{R}^{N \times N} \tag{2}$$

This attention map is used to aggregate the value vectors:

$$O = AV \in \mathbb{R}^{N \times d_e} \tag{3}$$

A final linear projection is applied to the output:

$$\hat{O} = OW^{\text{proj}}, \quad W^{\text{proj}} \in \mathbb{R}^{d_e \times d_e} \tag{4}$$

Each token is then independently passed through an MLP with two fully connected layers:

$$\text{MLP}(\mathbf{X}) = \sigma(\mathbf{X}W_1)W_2 \tag{5}$$

where $W_1 \in \mathbb{R}^{d_e \times 4d_e}$ and $W_2 \in \mathbb{R}^{4d_e \times d_e}$ are learned weight matrices and $\text{MLP}(\mathbf{X}) \in \mathbb{R}^{N \times d_e}$ is the output of the Transformer layer.

As for the computation complexity expressed in FLOPs, each module's complexity breaks down as follows:

$$\Phi_{\text{MHSA}}(N, d_e) = 4Nd_e^2 + 2N^2 d_e \tag{6}$$

$$\Phi_{\text{MLP}}(N, d_e) = 8Nd_e^2 \tag{7}$$

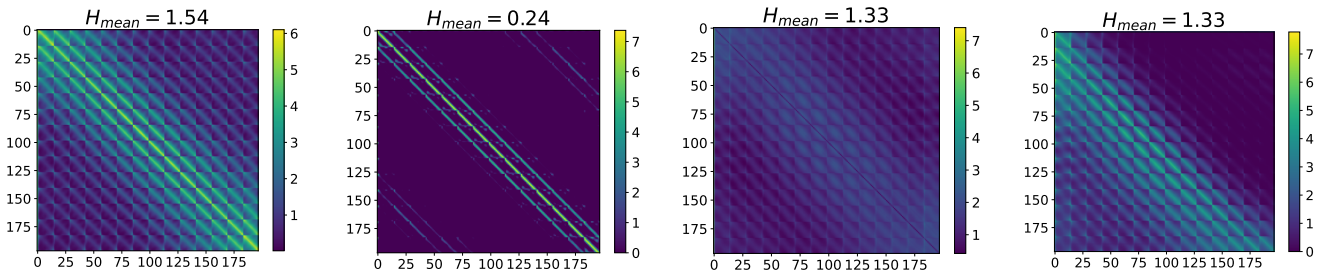
Finally, the total computational cost of a Transformer layer can be decomposed as follows:

$$\Phi_{\text{Layer}}(N, d_e) = \Phi_{\text{MHSA}}(N, d_e) + \Phi_{\text{MLP}}(N, d_e) \tag{8}$$

$$= 12Nd_e^2 + 2N^2 d_e \tag{9}$$

The expression in Eq (9) reveals that Transformers display quadratic complexity with respect to the sequence length N, as shown with the term $2N^2 d_e$. Besides, the MHSA mechanism requires the storage of the attention matrix $A \in \mathbb{R}^{N \times N}$, and the projection matrices (W_Q, W_K, W_V, W^{proj}) contribute by $O(d_e^2)$ parameters. Additionally, the MLP's weight matrices (W_1, W_2) introduce $O(d_e^2)$ parameters per layer. To alleviate this memory overhead, quantization emerges as a suitable optimization strategy.

As discussed in Section 1, attention weights are mainly concentrated to specific regions rather than uniformly distributed. This concentration creates redundancy, as many



(a) Layer 2 Head 2 (b) Layer 2 Head 3 (c) Layer 5 Head 1 (d) Layer 6 Head 3

Figure 3. Full entropy maps of a DeiT-Tiny computed on 5% of ImageNet-1K.

weights contribute minimally to the context of the input. Our approach is motivated by two hypotheses: (1) The redundancy in attention computations can be quantified by analyzing the entropy of attention weights across multiple input samples. (2) Attention weights that exhibit low entropy, demonstrating stable and predictable patterns across different inputs, can be frozen and quantized to low precision.

Specifically, for each attention head, we measure the entropy of its weight distributions over a dataset, which serves as an indicator of the weight stability. The details of this approach are formalized in the following section.

4. Methodology

4.1. Entropy: Information and uncertainty

Entropy, a fundamental concept in information theory, provides a robust metric to identify sensitive and useful parameters within a model. It has been widely used in various applications, including model quantization and regularization [25, 27], to optimize performance and reduce complexity. Entropy is defined as the amount of information contained in a probability distribution, representing the minimum number of bits required to encode the distribution without loss of information. Mathematically, the entropy $\mathcal{H}(X)$ of a random variable X with probability distribution $p(x)$ is given by:

$$\mathcal{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x), \text{ with } X \sim p(x) \tag{10}$$

Entropy can also be interpreted as a measure of uncertainty. In this context, information is inversely related to the predictability of an event. An event that is highly uncertain or surprising carries more information as it challenges our existing knowledge and expectations. This property makes entropy a valuable tool for assessing the informational content and sensitivity of the model parameters.

By combining attention scores with entropy, we can gain deeper insight into model dynamics and better discern important information from redundant information. In this way, we can better spot useful computation instead of useful parameters. As depicted above, the attention mechanism is a computationally expensive component and quadratically increases with the number of tokens N, the purpose of this paper is to find which computation in the attention mechanism is redundant and can be avoided using entropy.

4.2. Entropy Attention Maps

4.2.1. Definition

To better quantify the behavior of attention maps and their uncertainty, we compute the entropy of each of their weights. We take every weight as a random variable with some distribution p taking the values in $[0, 1]$.

We define the attention map of the layer l and the head h of an image m by:

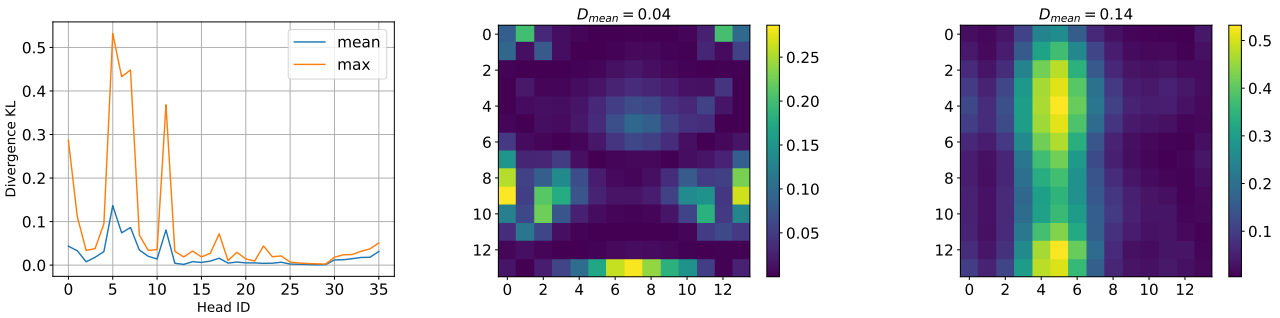
$$A^m_{l,h} = \text{softmax}\left(\frac{K^m_{l,h} Q^{l,m^T}_h}{\sqrt{d_l}}\right) \tag{11}$$

$A^m_{l,h} \in \mathbb{R}^{(N+1)\times(N+1)}$, since we add the Self-Attention of the CLS token. We estimate the distribution p of each attention weight i of this random variable with a histogram quantized in $b = 8$ bits. This way, we decompose the distribution into 256 values between 0 and 1 due to the softmax function.

$$p^i_{l,h}(k) = \frac{1}{M}\sum_m^M \left(A^m_{l,h}[i] \in \left[\frac{k}{2^b}, \frac{k+1}{2^b}\right[\right), \tag{12}$$
$$k \in \{0, 1, ..., 2^b - 1\}$$

M being the number of images in our dataset that we use to estimate the distribution, here we use 5% of ImageNet-1K. Thus, we can compute the entropy of every attention weight and quantize their uncertainty as follows:

$$\mathcal{H}_{l,h}[i] = -\sum_{k=0}^{2^b-1} p^i_{l,h}(k) \log_2(p^i_{l,h}(k)) \tag{13}$$



(a) (b) (c)

Figure 4. Visualization of KL divergence of attention weights distributions between the non-quantized model $p_{\text{fp32}}(x)$ and the 4-bit quantized model $p_q(x)$ on DeiT-Tiny, computed on 5% of ImageNet-1K. (a) Evolution of divergence of the CLS token over heads of the model. (b-c) Divergence maps on the CLS token of a DeiT-Tiny for Layer 1 Head 1, and Layer 2 Head 3 respectively.

where $\mathcal{H} \in \mathbb{R}^{L\times H\times(N+1)\times(N+1)}$. Through this process, we derive a matrix that encapsulates the uncertainty and redundancy for each weight in the attention maps.

4.2.2. Visualization

First, we visualize the entropy maps of the class token (CLS), specifically focusing on the first row of the entropy map, for the first six attention heads of the DeiT-Tiny model, as shown in Figure 2.

These visualizations reveal distinct behavioral differences between attention heads, indicating that some heads are more redundant than others. For instance, layer 2 head 1 (L2H1) exhibits entropy values near zero for most of its elements, thereby confirming our initial hypothesis. Importantly, this low entropy does not imply that the head is less important or can be discarded, rather, it suggests that its values are highly certain and stable. In fact, some heads, such as L2H1, demonstrate low entropy, indicating that their weights remain stable.

Examining a full entropic attention map in Figure 3 provides a clearer view of the uncertainty within an attention head. Once again, we observe distinct behaviors between two different heads. In particular, the matrix for L2H3 appears almost entirely empty of information, indicating high redundancy and predictability in many of its attention weights.

4.3. Impact of Quantization

EAM can also help us understand the impact of quantization on our model. If we see quantization as adding noise to the weights and activations, its impact should be detected in our entropy attention maps.

To better quantify the amount of noise introduced, we can use the KL divergence to measure the distance between two distributions:

$$D_{KL}(p_{\text{fp32}}||p_q) = \sum_{x\in\mathcal{X}} p_{\text{fp32}}(x)\log\frac{p_{\text{fp32}}(x)}{p_q(x)} \tag{14}$$

This formula allows us to compare the distribution of attention weights from the non-quantized model $p_{\text{fp32}}(x)$ with its quantized counterpart $p_q(x)$.

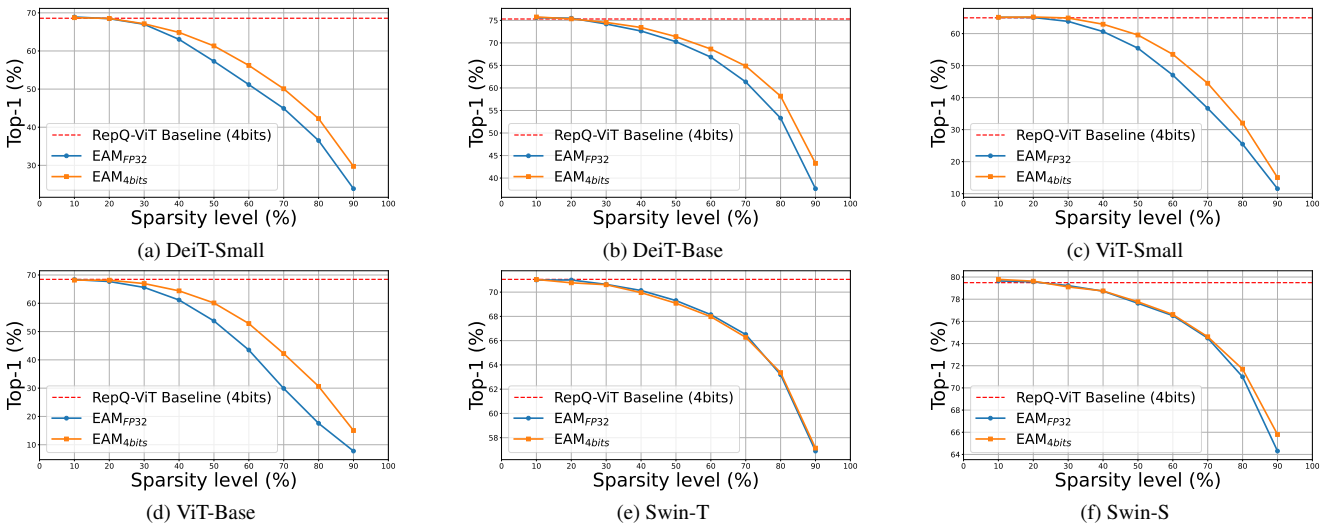
Figure 4 illustrates two examples of divergence maps on the CLS token for a DeiT-Tiny model with 4-bit quantized weights and activations, along with the evolution of the mean and maximum divergence within the model. We observe distinct behaviors among different heads; for instance, the fifth head (L2H3) exhibits higher divergence values compared to the first head. Additionally, quantization appears to have a more significant impact on the initial layers, gradually diminishing towards the end. This suggests that quantization affects each head uniquely, leading us to believe that EAM should be computed on the quantized version to more accurately represent the dynamics of the uncertainty introduced by quantization.

4.4. Attention weights fixing

As noted in previous sections, many attention weights are redundant and remain unchanged across different inputs. We can therefore fix these weights and replace them with their mean value. This approach allows us to reduce the computational cost of the model by bypassing a portion of the attention mechanism. A mask is applied to the attention map, where the values are defined as the $\tau\%$ lowest of the entropic attention map:

$$A^{\text{fix}}_{l,h} = A_{l,h}\otimes(H_{l,h} > \epsilon_\tau) + A^{\mu}_{l,h}\otimes(H_{l,h} < \epsilon_\tau), \tag{15}$$

$$A^{\mu}_{l,h} = \frac{1}{M}\sum_{m}^{M} A^{m}_{l,h}$$



(a) DeiT-Small (b) DeiT-Base (c) ViT-Small

(d) ViT-Base (e) Swin-T (f) Swin-S

Figure 5. Top-1 accuracy of EAM compared to the RepQ-ViT baseline for different sparsity levels across various ViT models.

In the following sections, we will refer to τ as the sparsity level, although in the literature this typically denotes matrices with $\tau\%$ zeros. In our context, the matrix $A_{l,h} \otimes (H_{l,h} > \epsilon_\tau)$ becomes sparse, containing zeros that will be replaced by the mean $A^{\mu}_{l,h}$.

5. Experiments

5.1. Set up

We conduct experiments on common ViT-based models, including ViT-Small, ViT-Base, DeiT-Tiny, DeiT-Small, DeiT-Base, Swin-Tiny, and Swin-Small using the timm package. For post-training quantization, we employ RepQ-ViT [18], which demonstrates excellent performance on image classification. We adhere to the original paper's configuration for quantization parameters and use ImageNet-1K to calibrate the entropy of attention maps.

Besides, we investigate the impact of attention sparsity on the accuracy of EAM across these ViT models, and we compare two variants of our model to the RepQ-ViT baseline. EAM$_{\text{FP32}}$, where model weights and activations are quantized to 4 bits while frozen attention map weights retain 32-bit precision, and EAM$_{\text{4bits}}$, where model weights, activations and frozen attention map weights are quantized to 4 bits. Finally, we conduct an ablation study comparing EAM with random fixing, where we use random selection and fixing of attention map weights instead of EAM.

5.2. Results

5.2.1. Impact of sparsity

Figure 5 illustrates the Top-1(%) accuracy of EAM on ImageNet-1K across varying sparsity levels for models quantized to 4-bit weights and activations. Two variants of EAM are compared: (1) EAM$_{FP32}$, derived from the full-precision (unquantized) model, and (2) EAM$_{4bits}$, computed using the quantized model via RepQ-ViT.

First, we observe that a sparsity level of up to 30% can be achieved across all models without significant accuracy degradation. By selectively targeting low-entropy attention weights, these can be frozen without compromising model performance. Second, the EAM$_{4bits}$ computation reduces the performance gap relative to the baseline, particularly at higher sparsity ratios (e.g., $> 50\%$) compared to EAM$_{FP32}$, this is true for all DeiT models.

Furthermore, at lower sparsity levels (10–20%), EAM$_{4bits}$ occasionally enhances accuracy compared to the RepQ-ViT baseline. For example, DeiT-Base achieves a Top-1 accuracy of 75.31% in the baseline configuration but improves to 75.71% at 10% sparsity and 75.64% at 20% sparsity. At intermediate sparsity levels (30-40%), the Top-1 accuracy drop is not significant compared to the gains in complexity.

Finally, we observe that Swin-based models behave somewhat differently and appear to be more robust to fixed attention weights. Specifically, they can tolerate up to 50% sparsity with less than a 2% drop in accuracy, compared to a 10% drop for DeiT-Small.

Table 1 shows the performance of EAM with 4-bit precision at $\tau = 10\%$ and $\tau = 20\%$ sparsity levels. In general, EAM leads to an increase in performance at these sparsity levels across all models except ViT-Base and Swin-Tiny, where the accuracy drop is minimal, at less than 0.3% and 0.12%, respectively. Although the improvement is modest, it is worth noting that fixing and quantizing the attention weights to 4 bits can contribute to enhanced performance. The following section will compare our entropy-based fix-

Method	Prec. (W/A)	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S
Full-Precision	32/32	81.39	84.54	72.21	79.85	81.80	81.30	83.23
RepQ-ViT	4/4	64.92	**68.46**	57.91	68.58	75.31	**70.67**	79.45
$\textbf{EAM}_{4bits}^{\tau=10\%}$	4/4	**65.09**	68.18	**58.03**	**68.74**	**75.71**	70.65	**79.79**
$\textbf{EAM}_{4bits}^{\tau=20\%}$	4/4	**65.19**	68.16	**58.02**	68.53	**75.64**	70.55	**79.63**

Table 1. Top-1 accuracy of EAM under 4-bit precision on various ViT models for 10% and 20% sparsity levels, compared to RepQ-ViT and Full-Precision baselines.

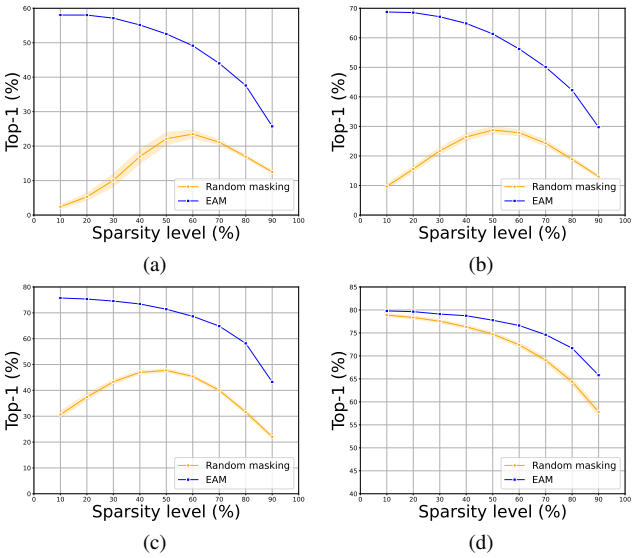


Figure 6. Ablation study of EAM against random fixing across sparsity levels. (a) DeiT-Tiny, (b) DeiT-Small, (c) DeiT-Base, (d) Swin-S.

ing of attention weights in EAM with random fixing.

5.2.2. Random fixing

To validate the effectiveness of our entropy-based fixing approach in EAM, we compare it against random fixing, i.e., randomly selected attention weights. Figure 6 illustrates the Top-1 accuracy versus the sparsity level of EAM against random fixing across DeiT-Tiny, DeiT-Small, DeiT-Base, and Swin-S, where our method consistently outperforms random fixing on all these models.

The main insights of this ablation study are two-fold. First, the discrepancy between the accuracy of EAM and random fixing is significant enough to validate our approach, especially on DeiT models, where Top-1 accuracy of random fixing collapses, keeping the gap to our model as low as 11.67% and as high as 55.60% on DeiT-Tiny. Although the gap is less pronounced on Swin-S, it remains statistically significant across sparsity levels, as the gap in Top-1 accuracy increases from 0.91% at 10% sparsity to 6.60% at 90% sparsity.

Second, DeiT is more sensitive to random fixing than Swin. This difference likely arises from the localized attention windows and hierarchical structure in Swin: While DeiT relies on global attention across all patches, Swin restricts interactions to local regions and refines features through downsampling. As a result, randomly fixing parts of attention maps in DeiT shows a higher impact on the subsequent layers compared to Swin, as the damage is confined to a single window, and later layers can recover lost information through merged features.

6. Conclusion

In this work, we introduced EAM, an entropy-driven approach to optimize Vision Transformers by analyzing and exploiting the information redundancy in attention heads. Our main insight is that low-entropy attention heads exhibit stable, predictable patterns across inputs, allowing us to fix and quantize them aggressively without compromising model performance. Through extensive experiments on ImageNet-1K with various ViT architectures, we demonstrated that our method reduces computational complexity and memory demands while maintaining accuracy. Specifically, EAM increases the Top-1 accuracy of the RepQ-ViT baseline while fixing 10% to 20% of the weights in attention maps and quantizing them to 4 bits. Furthermore, we achieved up to 40% sparsity in attention maps with negligible performance degradation. Finally, we validated the entropy-based fixing in EAM with an ablation study with random fixing, and showed that EAM outperforms random fixing on all the ViT models. In future work, we will extend our method to larger architectures, including Vision-Language Models (VLMs) and Large Language Models (LLMs), which process longer context sequences compared to ViTs. Given the high computational cost of MHSA, applying our method to these models is expected to yield greater computational savings, therefore, we find it interesting to validate these gains experimentally. In addition, we can further extend our work by enabling the model to retrain with attention weights fixed with EAM, aiming to minimize the loss in accuracy.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 2

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2023. 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] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[7] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022. 2

[8] Karim Haroun, Jean Martinet, Karim Ben Chehida, and Thibault Allenet. Leveraging local similarity for token merging in vision transformers. In *ICONIP 2024-31th International Conference on Neural Information Processing*, 2024. 2

[9] Karim Haroun, Thibault Allenet, Karim Ben Chehida, and Jean Martinet. Dynamic hierarchical token merging for vision transformers. In *VISAPP-2025-20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2025. 2

[10] Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–783, 2023. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2

[13] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1383–1392, 2024. 2

[14] Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In *International Conference on Learning Representations*, 2020. 2

[15] Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-vit: Accurate and fully quantized low-bit vision transformer. *Advances in neural information processing systems*, 35:34451–34463, 2022. 2

[16] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European conference on computer vision*, pages 154–170. Springer, 2022. 2, 3

[17] Zhexin Li, Tong Yang, Peisong Wang, and Jian Cheng. Q-ViT: Fully Differentiable Quantization for Vision Transformer, 2022. arXiv:2201.07703 [cs]. 3

[18] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17227–17236, 2023. 2, 3, 6

[19] Youwei Liang, Chongjian GE, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. EVit: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. 2

[20] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. *arXiv preprint arXiv:2111.13824*, 2021. 3

[21] Y. Liu, M. Gehrig, N. Messikommer, M. Cannici, and D. Scaramuzza. Revisiting token pruning for object detection and instance segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2646–2656, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[23] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2021. 2

[24] Zechun Liu, Kwang-Ting Cheng, Dong Huang, Eric P Xing, and Zhiqiang Shen. Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4942–4952, 2022. 2

[25] Lucas Maisonnave, Cyril Moineau, Olivier Bichler, and Fabrice Rastello. Applying maximum entropy principle on quantized neural networks correlates with high accuracy. In *AccML 2024-6th Workshop on Accelerated Machine Learning*, 2024. 4

[26] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018. 2

[27] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions, 2017. 4

[28] Mathilde Proust, Martyna Poreba, Michal Szczepanski, and Karim Haroun. Step: Supertoken and early-pruning for efficient semantic segmentation. In *VISIGRAPP 2025-20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2025. 2

[29] Navin Ranjan and Andreas Savakis. Mix-qvit: Mixed-precision vision transformer quantization driven by layer importance and quantization sensitivity. *arXiv preprint arXiv:2501.06357*, 2025. 2

[30] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 1

[31] Q. Tang, B. Zhang, J. Liu, F. Liu, and Y. Liu. Dynamic token pruning in plain vision transformers for semantic segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 777–786, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2

[32] Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhijun Tu, Kai Han, Hailin Hu, and Dacheng Tao. A survey on transformer compression. *arXiv preprint arXiv:2402.05964*, 2024. 1

[33] 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. 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[35] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8612–8620, 2019. 2

[36] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[37] Junrui Xiao, Zhikai Li, Lianwei Yang, and Qingyi Gu. Patch-wise mixed-precision quantization of vision transformer. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2023. 2

[38] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European conference on computer vision*, pages 191–207. Springer, 2022. 2, 3

[39] Edouard YVINEC, Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. Powerquant: Automorphism search for non-uniform quantization. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[40] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, and Yifan Liu. Segvit: Semantic segmentation with plain vision transformers. *NeurIPS*, 2022. 1

[41] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 2