Centered Kernel Alignment for efficient Vision Transformer quantization

Jose Lucas de Melo Costa, Cyril Moineau, Thibault Allenet, Inna Kucher

Centered Kernel Alignment for Efficient Vision Transformer Quantization

José Lucas De Melo Costa
CentraleSupélec[*]
Gif-sur-Yvette, France
jose-lucas.de-melo-costa@student-cs.fr

Cyril Moineau
Univ. Paris-Saclay, CEA LIST
F-91120, Palaiseau, France
cyril.moineau@cea.fr

Thibault Allenet
Univ. Paris-Saclay, CEA LIST
F-91120, Palaiseau, France
thibault.allenet@cea.fr

Inna Kucher
Univ. Paris-Saclay, CEA LIST
F-91120, Palaiseau, France
inna.kucher@cea.fr

***Abstract*—The rapidly evolving field of computer vision has witnessed a paradigm shift with the introduction of Transformer-based architectures, particularly Vision Transformers (ViTs). As these models expand in complexity, ensuring their efficient deployment on resource-limited devices becomes crucial. This paper proposes a solution for the model compression problem, emphasizing quantization, and highlights a notable gap in current methodologies: their need to consider outliers in the quantization process. We propose a distillation-guided quantization approach for ViTs, leveraging the Centered Kernel Alignment (CKA) similarity score. Empirical experiments are carried out on the DeiT architecture using the ImageNet dataset, with our CKA approach demonstrating promising results in retaining model intricacies during compression.**

***Index Terms*—Neural Networks, Model Compression, Quantization-Aware Training, Embedded Systems**

I. Introduction

Model compression has emerged as a significant challenge in deep learning and computer vision [8] in the context of embedded systems. The introduction of Transformers [24], originally for natural language processing (NLP), brought to the computer vision a shift with attention-based models like Vision Transformers (ViTs) [10] and other architectures [21, 18, 1]. As models become increasingly powerful, they often become complex, demanding more memory and computational resources. Such high-capacity models, while effective, become challenging to deploy on resource-constrained devices like mobile phones, edge IoT devices, or real-time applications [7]. Thus, the quest for efficient methods to reduce model size without compromising performance has become imperative.

To tackle this issue, knowledge distillation and quantization emerge as promising solutions: quantization, which involves representing continuous values (model parameters and activations) using fewer bits, and knowledge distillation, which creates a smaller model that learns from a larger one. Quantization can be done after training (Post-training quantization - PTQ) or integrated into the training process (Quantization-aware training - QAT), although the latter may require more time to train. Knowledge distillation, meanwhile, focuses on training a smaller model (the student) to mimic a large one (the teacher) by learning from its output or internal features. Combining this with quantization, models become more compact and run faster, particularly on devices tailored for such optimized models.

In this scenario, the natural language processing domain showcases the importance of outliers (e.g. that is, points that deviate slightly from others) in activations of the attention mechanism when compressing neural networks [5, 25, 6]. However, there needs to be more consideration for the role of these abnormal values when compressing models in the computer vision domain.

Inspired by this outlier problem outlined in the NLP context, our research introduces a quantization scheme that leverages the importance of outliers. We propose a quantization-aware training scheme that compares the internal structure of the quantized model with its full-precision counterpart. To do so, this comparison is made using the centered kernel alignment (CKA) metric. In this work, we opt for CKA due to its distinct sensitivity to outliers, its consistent performance in evaluating representational similarity compared to other metrics, and its resilience against translation as evidenced by findings in [9] and [14].

This work presents the ongoing experiments of compressing vision transformers (specifically the DeiT [21] model). As a result, the highlight of this investigation is the adaptation of the CKA metric to a quantization scheme, guiding the quantized model with the activations of the full-precision model. As this research is still a work in progress, the main contributions thus far include:

- Adaptation of the CKA similarity metric to knowledge distillation
- A distillation-guided quantization-aware training scheme for vision transformers

II. Backgroud and Related Work

Research into the quantization of Vision Transformers has gained traction, indicating its significance in contemporary machine learning. Before diving deep into the state-of-the-art methods, it's essential to understand the foundational concepts upon which they are built.

[*]This work was supported by a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program with the reference "ANR-23-DEGR-0001".

A. Baseline quantization

We briefly introduce the method for neural network quantization. A generic representation is given by

$$\mathcal{Q}_a(x, \alpha_x, \beta) = \Big\lfloor \texttt{clip}\Big(\frac{x-\beta}{\alpha_x}, Q_n^x, Q_p^x\Big)\Big\rceil \tag{1}$$

$$\hat{x} = \mathcal{Q}_a(x, \alpha_x, \beta) \times \alpha_x + \beta \tag{2}$$

$$\mathcal{Q}_{\mathbf{w}}(\mathbf{w}, \alpha_{\mathbf{w}}) = \Big\lfloor \texttt{clip}\Big(\frac{x}{\alpha_{\mathbf{w}}}, Q_n^{\mathbf{w}}, Q_p^{\mathbf{w}}\Big)\Big\rceil \tag{3}$$

$$\hat{\mathbf{w}} = \mathcal{Q}_{\mathbf{w}}(\mathbf{w}, \alpha_{\mathbf{w}}) \times \alpha_{\mathbf{w}} \tag{4}$$

where $\mathcal{Q}_a$ and $\mathcal{Q}_{\mathbf{w}}$ are quantization functions for the scalar activation x and weight vector $\mathbf{w}$, respectively. These functions utilize scaling factors, denoted by α_x and $\alpha_{\mathbf{w}}$, to adjust the range of quantization, and the term β is a zero-point bias. The `clip` function restricts its input to a specified range, and $\lfloor\cdot\rceil$ refers to rounding to the nearest integer. When quantizing the weights to b_1 bits and activations to b_2 bits (in this paper referred to as wb_1ab_2 quantization), we have $[Q_n^{\mathbf{w}} = -2^{b_1-1}, Q_p^{\mathbf{w}} = 2^{b_1-1}-1]$ and $[Q_n^x = -2^{b_2-1}, Q_p^x = 2^{b_2-1}-1]$, which are the negative and positive quantization limits for x and $\mathbf{w}$. Finally, $\hat{x}$ and $\hat{\mathbf{w}}$ represent the quantized outputs after applying the respective quantization functions.

A particularly influential method, the LSQ [11], introduced the idea of learning the quantization parameters (step size α_x, $\alpha_{\mathbf{w}}$ and bias β) during training. To achieve this, the straight through estimator (STE) [3] is used in the backward propagation, as

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial x} = \begin{cases} \frac{\partial \mathcal{L}}{\partial \hat{x}} & \text{if } x \in [-Q_n^x, Q_p^x] \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}}{\partial x}\frac{\partial x}{\partial \hat{\mathbf{w}}}\frac{\partial \hat{\mathbf{w}}}{\partial \mathbf{w}} = \begin{cases} \frac{\partial \mathcal{L}}{\partial x}\frac{\partial x}{\partial \hat{\mathbf{w}}} & \text{if } \mathbf{w} \in [-Q_n^{\mathbf{w}}, Q_p^{\mathbf{w}}] \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{L}$ is the loss function. This method was further explored and refined in LSQ+ [4] and Q-ViT [15].

B. Knowledge distillation

Knowledge distillation is the process where a smaller, often termed student, model is trained to replicate the behavior of a larger, teacher, model. The underlying idea is that the teacher, with its larger capacity, captures a more generalized representation of the data, and the student can benefit from this knowledge without bearing the computational burdens of the teacher.

The usual practice is using distillation through attention as described in [22]. In the case of hard distillation, the output of the teacher network is first evaluated as a predicted class y_t, as in:

$$y_t = \underset{c}{\operatorname{argmax}}\ z_t(c) \tag{5}$$

where z_t are the logits output for each class c. Then the usual cross-entropy loss can be calculated both for the correct label y and the teacher hard decision y_t, as in:

$$\mathcal{L}_{\text{dist}} = \frac{1}{2}\mathcal{L}_{\text{CE}}(\sigma(\mathrm{z}_s), y) + \frac{1}{2}\mathcal{L}_{\text{CE}}(\sigma(\mathrm{z}_s), y_t) \tag{6}$$

where $\mathcal{L}_{\text{CE}}$ stands for the cross-entropy loss, z_s are the logits of the student network and σ is the softmax function.

Distillation through attention, as described in [22], has become a common practice in the field. Building on these foundational concepts, there have been notable advancements in the field:

- The Q-ViT study [15] highlighted the challenges of information distortion within the attention mechanism. They introduced the Information Rectification Module (IRM) and utilized a distribution-guided distillation strategy.
- Mixed precision techniques have grown in appeal. For instance, Mixed-Q-ViT [16] proposed an approach for learning quantization scales and bit-widths concurrently, driven by the classification loss.
- Post-training quantization, as seen in FQ-ViT [17], has also made significant progress. The study presented innovative methods like Power-of-Two Factor (PTF) and Log-Int-Softmax (LIS) to address certain complexities.

III. Methodology

A. Centered kernel alignment

The centered kernel alignment (CKA) shows the similarity between pairs of activation matrices, as presented in [19]. Given CKA's inherent sensitivity to invertible linear transformations [9], our hypothesis is that it might become especially effective in spotlighting outliers. Also, CKA is notably superior in evaluating potential representations more consistently than other metrics, such as canonical correlation analysis and cosine similarity [14].

To formalize this idea, we first define the following terms:

$$S_L \equiv \frac{1}{k}\sum_{i=1}^{k} \text{HSIC}(\mathbf{L}_i, \mathbf{L}_i), \tag{7}$$

$$S_M \equiv \frac{1}{k}\sum_{i=1}^{k} \text{HSIC}(\mathbf{M}_i, \mathbf{M}_i), \tag{8}$$

$$S_{LM} \equiv \frac{1}{k}\sum_{i=1}^{k} \text{HSIC}(\mathbf{L}_i, \mathbf{M}_i), \tag{9}$$

where:

- $\mathbf{L}_i$ and $\mathbf{M}_i$ are linear kernel matrices derived from activation matrices in the i^{th} minibatch from a total of k minibatches comprising n samples [19].
- HSIC, the Hilbert-Schmidt Independence Criterion, is a method to measure statistical dependence between two sets of random variables [12, 13].

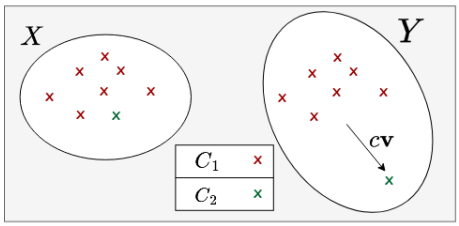

Fig. 1: Visual representation of the transformation. This representation was inspired by [9].

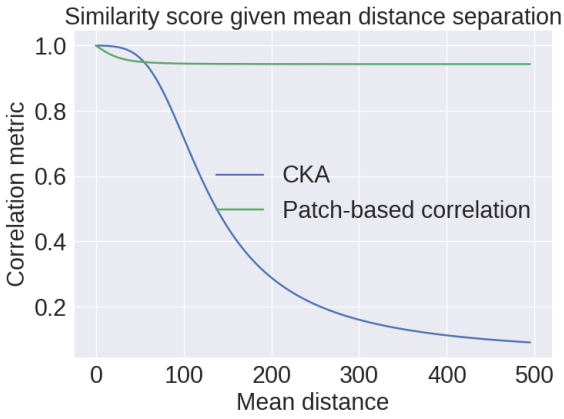

Fig. 2: Comparing the outlier sensitivity between the CKA and the correlation metric used in [15, 23].

With these definitions in place, the CKA for a minibatch is given by:

$$\text{CKA}_{\text{minibatch}} = \frac{S_{LM}}{\sqrt{S_L \cdot S_M}}. \tag{10}$$

We designed an experiment to assess an outlier sensitivity: a dataset is created from a Gaussian distribution, 2% of its points are shifted to simulate outliers, and similarity metrics are compared between original and shifted sets, as in Figure 1. Formally, let $X \in \mathbb{R}^{N \times d}$ be a set of N d-dimensional points with subsets $C_1 \subset X$ and $C_2 = X \setminus C_1$. A new set Y can be constructed as $Y = \{\mathbf{x} \,|\, \mathbf{x} \in C_1\} \cup \{\mathbf{x} + c\mathbf{v} \,|\, \mathbf{x} \in C_2\}$ where $\mathbf{v}$ is a d-dimensional unit vector and c is a scalar for translation distance.

It is possible to see, as presented in Figure 2, that when few data points are selected (e.g., $\frac{|C_2|}{|C_1|} \approx 0.02$, with $|A|$ the size of the set A), the CKA similarity is much more sensitive than the patch-based correlation metric used in [15, 23]. Considering matrices X and Y, the patch-based correlation is given by:

$$corr(X, Y) = \left|\left| \frac{XX^T}{||XX^T||_2} - \frac{YY^T}{||YY^T||_2} \right|\right|_F \tag{11}$$

where $||.||_2$ is the L_2 row-wise norm and $||.||_F$ is the Frobenius norm.

B. Proposed approach: CKA-ViT quantization

Now, we introduce the proposed apprach, which is a quantization-aware training scheme that uses the CKA to leverage the information of outliers in activations of the attention mechanism. The attention block takes three matrices: query $\mathbf{Q} \in \mathbb{R}^{N \times d_h}$, key $\mathbf{K} \in \mathbb{R}^{N \times d_h}$, and value $\mathbf{V} \in \mathbb{R}^{N \times d_v}$ with dimensions defined by d_h (for $\mathbf{Q}$ and $\mathbf{K}$), d_v (for $\mathbf{V}$), and N (input patches) [10]. The output is:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}}\right)\mathbf{V} \tag{12}$$

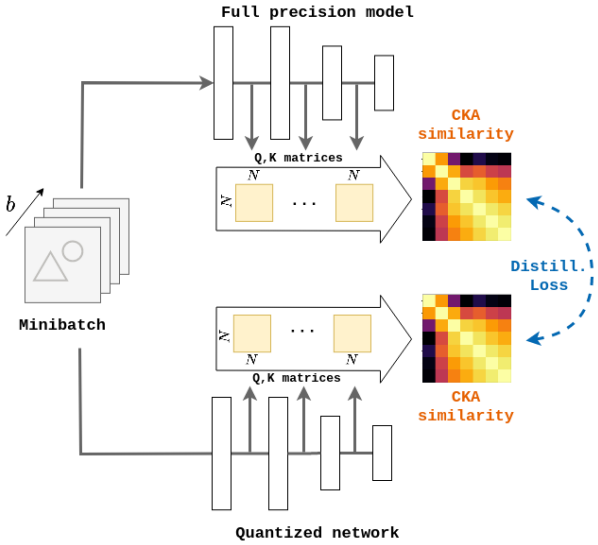

Fig. 3: Architecture of the proposed distillation loss. The centered Kernel Alignment (CKA) aids in gauging the similarity between representations. This diagram was inspired by [23].

To adapt the CKA metric, the queries and keys are gathered, and this information is summarized in the similarity matrices, as presented in Figure 3. For instance, the similarity matrix for the queries is defined as $(\mathbf{Q}^s_{CKA})_{ij} = CKA_{\text{minibatch}}(\mathbf{Q}^s_i, \mathbf{Q}^s_j)$, where $\mathbf{Q}^s_i$ and $\mathbf{Q}^s_j$ are the queries of the quantized model for the i-th and j-th blocks, respectively. The same is done for the keys $\mathbf{K}^s_{CKA}$. The representation loss, which is the extra loss that represents the discrepancies between the query and key internal representation of the models, is defined as:

$$\mathcal{L}_{CKA} = ||\mathbf{Q}^s_{CKA} - \mathbf{Q}^t_{CKA}||_2 + ||\mathbf{K}^s_{CKA} - \mathbf{K}^t_{CKA}||_2 \tag{13}$$

with $\mathbf{Q}^t_{CKA}$ and $\mathbf{K}^t_{CKA}$ the similarity matrices of the full-precision model. The CKA loss is added to the cross-entropy losses using a balancing factor γ. The chosen knowledge distillation scheme is presented in [21]. As such, the final loss, as presented, is given by:

$$\mathcal{L} = \gamma\mathcal{L}_{dist} + (1 - \gamma)\mathcal{L}_{CKA} \tag{14}$$

where $\mathcal{L}_{dist}$ is the distillation loss as in Equation 6. This loss ensures that not only are the final predictions of the student and

teacher aligned, but also their intermediate learned features, thus extracting a richer representation from the teacher.

IV. Experiments

A. Dataset and training details

All results are obtained using the ImageNet-1K dataset [20], a dataset composed of 1.2 million images and 1000 classes. The images are resized to 224×224 pixels, and normalized using the mean and standard deviation of the ImageNet dataset. The dataset is split into three parts: the training set, the validation set and the test set.

The training process used an NVIDIA A100 GPU, complemented by AMD EPYC 7502 processors. The CKA-ViT model, which we proposed, adopted a cosine learning rate scheduler that started at a rate of 3.75e-05, without any warmup phase. For optimizing the model's weights throughout the backpropagation, we opted for the Lamb optimizer [26]. This entire training was structured to run across a maximum of 50 epochs. Furthermore, for quantization-aware training, we integrated data augmentation strategies inspired by the methods presented in [22].

B. Results

Table I shows the effects of different quantization techniques on the DeiT Tiny model's size, FLOPs, and Top-1 accuracy.

Method	#Bits	Size$_{(MB)}$	FLOPS$_{(M)}$	Top-1 Acc. (%)
Full precision	w32a32	20.00	1382	72.21
FQ-ViT	w8a8	5.00	345	71.61
LSQ	w4a4	2.50	172	73.10
Q-ViT	w4a4	2.50	172	74.30
Mixed-Q-ViT	*w4a4	2.50	172	72.80
CKA-ViT (Ours)	w4a4	2.50	172	**75.39**
LSQ	w2a2	1.25	86	46.44
CKA-ViT (Ours)	w2a2	1.25	86	**51.31**

TABLE I: Quantization outcomes for the DeiT Tiny model across different precision settings. The nomenclature wb_1ab_2 refers to the bit precision b_1 of the weights and b_2 of the activations. * This method uses mixed precision, bounding to 4-bit quantization.

The study [21] found that the full precision model had 72.21% accuracy at 32 bits, dropping slightly to 71.61% with 8-bit FQ-ViT quantization. Intriguingly, some QAT methods even outperformed full precision in 4-bit quantization, likely due to extended training and regularization effects of quantization [2]. As shown in Figure 4, the typical similarity metric (in green) exhibits extreme values in the query matrix when compared with the CKA (blue). This confirms that the CKA-ViT method successfully enhances information extraction in quantization, outperforming other techniques.

To evaluate the effect of the CKA quantization, the infinite norm ($||\mathbf{x}||_\infty = \max_i |x_i|$) of the values from the query matrices were calculated for the CKA-ViT, the full precision model and the commonly used correlation [15, 23] as baseline. Given that outliers in this context are values that are exceptionally low or exceedingly high, the use of the infinite norm is justified due to its inherent capability to capture these extreme absolute values. In the context of quantization, outliers can induce significant distortions, especially if not properly addressed during the quantization process. And by using the infinite norm, we aimed to examine these extreme values.

It is possible to see that the usual similarity metric (represented in green) presents more extreme values in the query matrix, which will later influence the attention scores, as presented in Equation 12. This result is in line with the expected observations and justifies the use of the CKA metric.

Fig. 4: Infinite norm ($||\mathbf{x}||_\infty = \max_i |x_i|$) of the values of the query matrix, comparing the CKA-ViT (blue), the baseline correlation used in the Q-ViT approach (in green) and the full precision model (in red).

V. Conclusion and Future Directions

This study embarked on an exploration of quantization techniques, with a particular focus on the DeiT Tiny model, demonstrating the potential of strategic quantization in neural network optimization. Our proposed CKA-ViT model, a quantization-aware training method for vision transformers using the CKA metric, was examined using the ImageNet-1K dataset. The results, as highlighted in Table I, showed that our model could achieve significant model size reductions, ranging from 4x to 8x, with only a minimal compromise in accuracy. These findings not only highlight the efficacy of our CKA-ViT model but also bring into focus the significance of outliers in the quantization process. This is particularly relevant in light of our hypothesis.

The study's results indicate that our CKA-ViT model consistently outperformed other quantization methods across different precision settings, emphasizing the advantages of strategic quantization. However, it is crucial to note the time inefficiency and the computational complexity of the CKA metric in our approach. These aspects suggest potential areas for improvement and developing an approximation for the CKA could potentially alleviate this.

Furthermore, the experiments' scope was limited to the DeiT-Tiny architecture. For a comprehensive understanding and to fully harness the benefits of this method, future research should broaden its focus to encompass other ViT architectures,

particularly the larger models, and also consider alternative similarity metrics to the CKA.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Herve Jegou. Xcit: Cross-covariance image transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20014–20027. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a655fbe4b8d7439994aa37ddad80de56-Paper.pdf.

[2] MohammadHossein AskariHemmat, Reyhane Askari Hemmat, Alex Hoffman, Ivan Lazarevich, Ehsan Saboori, Olivier Mastropietro, Sudhakar Sah, Yvon Savaria, and Jean-Pierre David. QReg: On Regularization Effects of Quantization, June 2022. URL http://arxiv.org/abs/2206.12372. arXiv:2206.12372 [cs].

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation, August 2013. URL http://arxiv.org/abs/1308.3432. arXiv:1308.3432 [cs].

[4] Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2978–2985, 2020. doi: 10.1109/CVPRW50498.2020.00356.

[5] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/2021.emnlp-main.627.

[6] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable Transformers: Removing Outliers by Helping Attention Heads Do Nothing, June 2023. URL http://arxiv.org/abs/2306.12929. arXiv:2306.12929 [cs].

[7] Meng Chen, Jun Gao, and Wuxin Yu. Lightweight and Optimization Acceleration Methods for Vision Transformer: A Review. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2154–2160, October 2022. doi: 10.1109/ITSC55140.2022.9921989.

[8] Tejalal Choudhary, Vipul Mishra, Anurag Goswami, and Jagannathan Sarangapani. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53(7):5113–5155, October 2020. ISSN 1573-7462. doi: 10.1007/s10462-020-09816-7. URL https://doi.org/10.1007/s10462-020-09816-7.

[9] MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=8HRvyxc606.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[11] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=rkgO66VKDS.

[12] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 63–77, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31696-1. doi: 10.1007/11564089_7.

[13] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 585–592, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

[14] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

[15] Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-vit: Accurate and fully quantized low-bit vision transformer. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=fU-m9kQe0ke.

[16] Zhexin Li, Tong Yang, Peisong Wang, and Jian Cheng. Q-ViT: Fully Differentiable Quantization for Vision Transformer, September 2022. URL http://arxiv.org/abs/2201.07703. arXiv:2201.07703 [cs].

[17] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint*

Conference on Artificial Intelligence, IJCAI-22, pages 1173–1179. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/164. URL https://doi.org/10.24963/ijcai.2022/164. Main Track.

[18] 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

[19] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KJNcAkY8tY4.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. URL http://arxiv.org/abs/1409.0575. arXiv:1409.0575 [cs].

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

[22] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, page 516–533, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_30. URL https://doi.org/10.1007/978-3-031-20053-3_30.

[23] Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1365–1374, 2019. doi: 10.1109/ICCV.2019.00145.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010. Curran Associates Inc., 2017. ISBN 9781510860964.

[25] Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: Pushing the limit of low-bit transformer language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17402–17414. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f6db140de9c9f111b12ef8a216320a9-Paper-Conference.pdf.

[26] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Syx4wnEtvH.