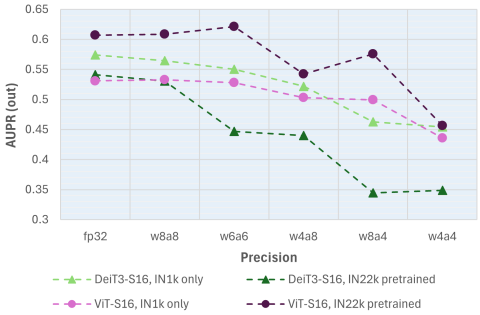
Understanding vision transformer quantization robustness through the lens of out-of-distribution detection

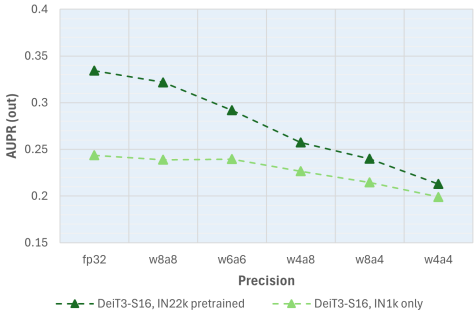
Joey Kuang, Alexander Wong
Vision and Image Processing Group, Department of Systems Design Engineering
University of Waterloo, Ontario, Canada
{jjykuang,a28wong}@uwaterloo.ca

Abstract

Vision transformers have shown remarkable performance in vision tasks, but enabling them for accessible and real-time use is still challenging. Quantization reduces memory and inference costs at the risk of performance loss. Strides have been made to mitigate low precision issues mainly by understanding in-distribution (ID) task behaviour, but the attention mechanism may provide insight on quantization attributes by exploring out-of-distribution (OOD) situations. We investigate the behaviour of quantized small-variant popular vision transformers (DeiT, DeiT3, and ViT) on common OOD datasets. ID analyses show the initial instabilities of 4-bit models, particularly of those trained on the larger ImageNet-22k, as the strongest FP32 model, DeiT3, sharply drops 17% from quantization error to be one of the weakest 4-bit models. While ViT shows reasonable quantization robustness for ID calibration, OOD detection reveals more: ViT and DeiT3 pretrained on ImageNet-22k respectively experienced a 15.0% and 19.2% average quantization delta in AUPR-out between full precision to 4-bit while their ImageNet-1k-only counterparts experienced a 9.5% and 12.0% delta. Overall, our results suggest pretraining on large scale datasets may hinder low-bit quantization robustness in OOD detection and that data augmentation may be a more beneficial option.



(a) AUPR-out averaged across five OOD datasets (all listed in Sec. 2, except for ImageNet-O. Average random chance level is 0.175.



(b) AUPR-out on ImageNet-O. Random chance level is 0.167.

Figure 1. Higher performing models assign OOD examples with higher anomaly scores. See Sec. 2 for a description of the AUPR-out as an OOD detection measure. Models pretrained on the larger dataset are more sensitive to quantization error in OOD detection.

1. Introduction

In recent years, vision transformers (ViTs) have risen to be top performers in large-scale vision tasks and are becoming more widely considered as a new standard in deep computer vision. They rely on the necessarily quadratic self-attention mechanism [8], employing multiple heads of attention in one block, with several blocks in its large backbone. This makes ViTs poor candidates for edge deployment. However, model compression techniques such as quantization may enable ViTs for more powerful real-time use cases.

Quantization compresses a model by reducing the precision required to represent the parameters, reducing both memory overhead and computational complexity. Lowering the precision introduces noise that may cause performance deltas. There are two general frameworks of quantization: quantization-aware training (QAT) and post-training quantization (PTQ). QAT involves fully re-training the model with objectives that minimize the quantized model error. This is often resource-consuming and challenging to

	Top-1 accuracy (%)						NLL					
	FP32	W8A8	W6A6	W4A8	W8A4	W4A4	FP32	W8A8	W6A6	W4A8	W8A4	W4A4
DeiT3-S16, IN22k pretrained	82.700	82.594	81.380	73.491	76.287	65.288	0.733	0.819	1.002	1.171	1.325	1.662
DeiT3-S16, IN1k only	81.302	81.212	80.144	74.768	74.052	66.873	0.894	0.897	0.974	1.334	1.440	1.927
ViT-S16, IN22k pretrained	81.337	81.151	80.265	75.484	73.173	64.812	0.673	0.682	0.726	0.926	1.057	1.486
DeiT-S16, IN1k only	79.786	79.714	78.829	74.831	73.367	67.933	0.883	0.888	0.945	1.148	1.242	1.512

Table 1. Quantization performance on ImageNet-1k (in-distribution). Each datapoint is an average across three seeds. All models have the same architecture and only differentiate by training method and dataset, but demonstrate varying levels of robustness to lower precision.

	VIT-S16, IN22k [25]	DeiT-S16, IN1k [26]	DeiT3-S16, IN1k [27]	Deit3-S16, IN22k [27]
Batch size	4096	1024	2048	2048
Num epochs (PT/FT)	300/30	300	400	90/50
Optimizer	AdamW	AdamW	LAMB	LAMB
Weight decay	0.1	0.05	0.02	0.02
Label smoothing ϵ	0.1	0.1	✗	0.1
Gradient Clip	Norm, 1.0	✗	Norm, 1.0	Norm, 1.0
RRC	✓	✓	✓	✗
RandAugment ($N/M/M_{std}$)	2/15/0.0	2/9/0.5	✗	✗
3 Augment	✗	✗	✓	✓
LayerScale	✗	✗	1e-4	1e-4
Mixup alpha	0.2	0.8	0.8	✗
Cutmix alpha	✗	1.0	1.0	1.0
Erasing prob.	✗	0.25	✗	✗
ColorJitter	✗	✗	0.3	0.3

Table 2. Summary of training procedures for the models evaluated. For ViT by [25], note that the model trained on ImageNet-1k only is also trained for 300 epochs and the configuration was taken from the specific model checkpoint[1]. We show this table to highlight differences in training methods that may contribute to our observations in quantized OOD performance.

configure [14]. PTQ is a less precise, but more practical solution, requiring only a small calibration dataset to determine quantization parameters [20]. While convolutional neural networks (CNNs) have enjoyed success from techniques developed in this field, ViTs could not see translated benefits. Early works exploring ViT quantization have observed varying complex distributions in LayerNorm, Softmax, and GeLU activations (the former two being part of the attention block), requiring the proposal of new PTQ schemes [7, 16, 17, 32].

While there is commendable progress in closing the quantization gap for accuracy, we also wonder if the same can be said for out-of-distribution (OOD) tasks. If we consider ViTs for deployment, then real-time usage will challenge models with unseen, noisy inputs. In their full precision form, the features produced by attention blocks have shown to be valuable in general robustness to distribution shifts, citing lack of inductive bias allowing for richer representation learning [4, 9, 19, 24]. However, Pinto *et al.* [22] discuss the role of pretraining actually being more significant than the inductive bias when comparing ViTs to CNNs. In any case, quantization integrity in the attention block is ostensibly crucial to its performance for any distribution.

We investigate whether OOD-related tasks can add insight on common vision transformer robustness to quantization. We test quantized models on six common OOD datasets and evaluate their ability to distinguish OOD samples from in-distribution (ID) ones. Results shown in Fig. 1 suggests that large-scale pretraining hinders low-bit quantization robustness and this is made apparent in OOD detection. This leads us to visualize attention to see whether our quantitative results align with our expectations in quantized attention integrity. Our findings suggest that data augmentation is beneficial for OOD tasks even in lower precision.

2. Methodology

2.1. Motivation

The vision transformer architecture embeds image patches which are fed into blocks, each composed of a multi-head

[1] https://storage.googleapis.com/vit_models/augreg/[S_16-i21k-300ep-lr_0.001-aug_medium2-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.03-res_224.npz,S_16-i1k-300ep-lr_0.001-aug_medium2-wd_0.1-do_0.0-sd_0.0.npz]

self-attention (MHSA) module and a multi-layer perceptron module. The MHSA projects these tokenized patches into three representations (query, key, value) and applies Softmax to represent attention scores in $[0, 1]$. These attention scores aim to inform the model of relevant inter-patch correlations; preserving attention features in quantization is therefore crucial in vision tasks. We propose that this goes beyond maintaining the performance in classifying ID samples, but also preserving learned features that help the model distinguish OOD samples from ID.

We would like to probe the quantized vision transformer's behaviour in both an ID and OOD setting. If the behaviour is aligned between distribution shifts, then the current paradigm of benchmarking quantization performance can likely continue as is. Otherwise, we may gain new insight on how we can approach vision transformer quantization. To narrow down the model behaviour in OOD contexts, we specifically look at models that differ only by training protocol; previous works have discussed the value of certain training techniques in model robustness [10, 12].

We start with a brief analysis on an ID task using classification accuracy and negative-log likelihood (NLL), common metrics that practictioners use to benchmark quantization robustness. The latter is a proper score and strong indicator of model calibration, which is relevant for robustness to different distributions. Since we are hypothesizing that quantization of the attention mechanism can lead to a deterioration in learned correlations, we should ask how well can the model distinguish a sample as OOD when it correctly identifies it as so vs. when it incorrectly detects as so. We use maximum softmax probability, suggested by Hendrycks *et al*. [11], to obtain the Area Under Precision-Recall (AUPR) scores for OOD samples (OOD as the 'positive' class and ID as the 'negative'). The following section describes the experiment setup in detail.

2.2. Setup

We investigate small-variant vision transformers of the same architecture, either pretrained on ImageNet-22k (IN22k) [6] or ImageNet-1k (IN1k) [23]: ViT [25], DeiT [26], and DeiT3 [27]. These size variants are commonly studied in other quantization works, and are more relevant in the context of edge devices (albeit still massive for edge). Model weights are obtained from Timm library[2] and from Google Research's cloud storage.

We quantize all networks using the open-sourced RepQ-ViT [15, 16], a strong PTQ method for ViTs. We follow the original paper's quantization configuration. Seeds 0, 13, 37 are used to generate a robust set of results. For evaluation on ID performance, we use ImageNet-1k. For OOD evaluation, we use six commonly tested datasets: ImageNet-O [13], Textures [3], Places365 [33], SUN [30], iNaturalist [28], and OpenImage-O [29]. The baseline ID analysis reports Top-1 accuracy and NLL for insight into quantization errors related to task performance and calibration. We report AUPR-out [11] scores to evaluate OOD robustness, using ImageNet-1k validation set as the ID comparison set. For attention visualization, we use attention rollout and follow Gildenblat's suggestion on discarding lowest pixels (ratio of 0.9) and taking the maximum value among the heads for best visualization [1, 2].

[2]https://github.com/huggingface/pytorch-image-models/tree/main/timm

3. Key findings

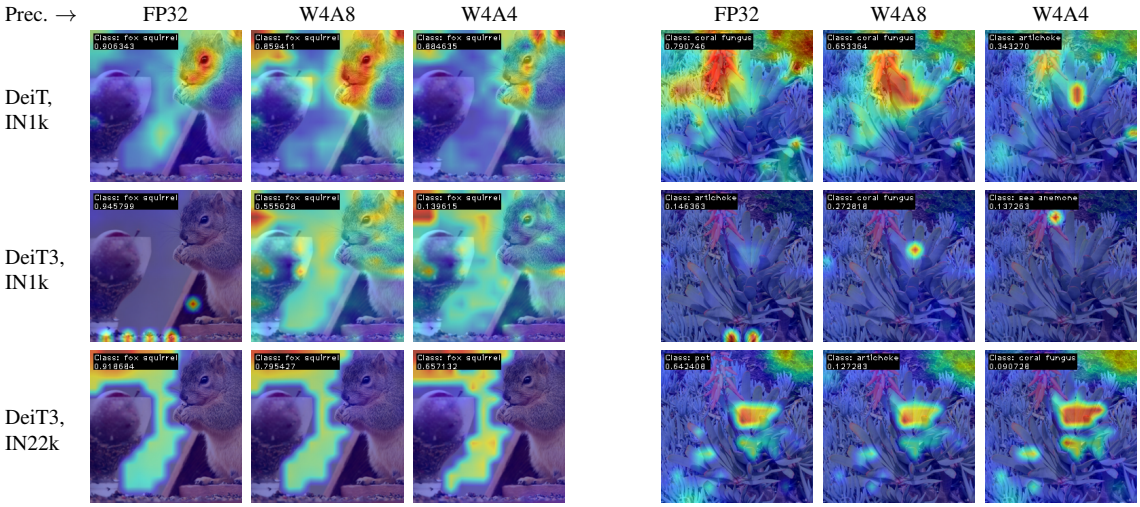
3.1. Vision transformers have unpredictable levels of robustness to 4-bit quantization

ViT-likes are fairly robust to 8- and 6-bit quantization under the RepQ-ViT framework. Once the model faces 4-bit quantization, we start to see instability in the highest performing models. Tab. 1 shows the strongest FP32 model, DeiT3-S16 (IN22k pretrained), steeply drop over 17% in accuracy at full 4-bit. ViT-S16 (IN22k pretrained) suffers a 16.5% drop, while the two IN1k-trained models are better able to retain classification performance at 4-bits (DeiT3 at 14.4% and DeiT at 11.9%). The differences in pretraining scale do not explain every result, as both DeiT3 models suffer a sharp increase in NLL. Meanwhile, ViT-S16 is consistently the most calibrated model at every precision, and DeiT-S shows minimal quantization error again.

Since all models have the same architecture, we can hone in on training procedure. We highlight the training methods between the four models in Tab. 2. Of note for DeiT is the lack of gradient clipping and the large amount of data augmentation applied. Gradient clipping stabilizes the optimization process in training, nearly guaranteeing convergence [18, 21]. If this technique lends to poor 4-bit stability, it implies that optimal model parameters in their trained data range (FP32) have severely high variance. More plausible is the idea that DeiT's aggressive use of data augmentation have flattened their loss landscape, an observation explored by [10], which seems to have translated in quantization. The remainder of the investigation focuses on the two pairs of IN1k-trained and IN22k-pretrained models by [25] and [27] to further scrutinize effects of pretraining scale and other training decisions.

3.2. Vision transformers pretrained on ImageNet-22k are less robust to quantization in OOD

We now analyze ViT and DeiT3 in OOD scenarios. Although pretraining was not a performance indicator for the ID case, we see a more consistent pattern in quantization stability for OOD tests. The IN22k-pretrained models sharply drop in their ability to distinguish OOD samples from ID, regardless of their capabilities at FP32. In Fig.



(a) Attention rollout maps for an ID sample (ImageNet-1k).

(b) Attention rollout maps for an OOD sample (iNaturalist).

Figure 2. DeiT [26] produces interpretable attention maps even at lower precision. We observe outlier artifacts similar to findings in [5] in DeiT3 features. Despite poor AUPR robustness to 4-bit quantization, attention maps produced by DeiT3-S pretrained on IN22k seem stable. Attention maps do not appear to vary greatly under OOD data. Predicted class label and probability are labeled in each image.

1(a), the 3.3% gap between the two FP32-DeiT3 models widens to 10.5% at 4-bit, and the IN1k variant consistently outperforms the IN22k-pretrained variant. We find a similar behaviour in stability on ImageNet-O. This is significant since the IN22k-pretrained model has actually seen this data (ImageNet-O is constructed from a subset of IN22k), so we expect a better ability to distinguish it as OOD. Tab. 2 shows the additional augmentations of RRC and Mixup on the side of the IN1k-trained model. This further supports the benefits of data augmentations for loss landscapes translating under quantized settings.

We also see that a 7.6% AUPR-out gap between the two FP32-ViT models diminishes to 2.1% at 4-bit. In this case, the only difference in training is the data seen and total time training, again suggesting that large-scale pretraining may have adverse effects to OOD data in the low-bit regime.

3.3. High-norm outlier tokens in attention maps may be too dominant in lower precision

To illustrate the effects of quantization on vision transformers, we visualize the attention attribution at each precision for the DeiT and DeiT3 models in Fig. 2. DeiT3 features contain a lot of artifact patches (hot spots in irrelevant areas) resembling the phenomenon of high-norm outlier tokens discussed by [5]. Darcet *et al*. [5] investigated this phenomenon and suggested that pretraining at a large scale for longer durations and larger model sizes can lead to forming outlier tokens. They also find that these outlier tokens contain global information. Given that we use the Small variant, the training duration of DeiT3 is a plausible contributor in learning these tokens.

Interestingly, DeiT3-S-IN22k attention maps appear to be robust at 4-bits; in the IN1k-trained counterpart maps, the outlier tokens "move" across the image, suggesting sensitivity to 4-bits. While this aligns with the quantization error in ID calibration, the changes in outlier attributes did not hinder OOD detection capabilities. Yellapragada *et al*. [31] show that the learned representations of high-norm tokens distilled into registers do not generalize well in OOD. The IN22k-pretrained model appears to have retained the global information of these outlier tokens at lower bits, but it is possible that other local representations may have been saturated. Meanwhile, the outlier norm in the IN1k model may not have been severe, which causes quantization error in the global information at lower bits (adapting the range to a lower magnitude), but other information useful for anomaly detection is better preserved.

4. Conclusion

Efforts to improve vision transformer performance in the quantized regime have not necessarily translated in OOD scenarios. We present findings that suggest that pretraining and data augmentation play a non-trivial role in dictating model stability in low-bit quantization. We also give an initial look into how high-norm tokens holding global information behave under quantization and how that affects anomaly detection. Future works include further investigation into how the representations learned by high-norm tokens and other patch tokens behave at lower precision, and into the value of data augmentation and large-scale pretraining for quantization in OOD.

References

[1] Exploring explainability for vision transformers. 3

[2] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197. Association for Computational Linguistics. 3

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. pages 3606–3613. 3

[4] Luca Cultrera, Lorenzo Seidenari, and Alberto Del Bimbo. Leveraging visual attention for out-of-distribution detection. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4449–4458. IEEE. 2

[5] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. ISSN: 1063-6919. 3

[7] Yifu Ding, Haotong Qin, Qinghua Yan, Zhenhua Chai, Junjie Liu, Xiaolin Wei, and Xianglong Liu. Towards accurate post-training quantization for vision transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5380–5388. Association for Computing Machinery. 2

[8] Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, pages 597–619. PMLR, 2023. 1

[9] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 7068–7081. Curran Associates, Inc. 2

[10] Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. 3

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 3

[12] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2712–2721. PMLR, . ISSN: 2640-3498. 3

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples, . 3

[14] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. 2

[15] Zhikai Li, Xuewen Liu, Jing Zhang, and Qingyi Gu. RepQuant: Towards accurate post-training quantization of large transformer models via scale reparameterization, . 3

[16] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. RepQ-ViT: Scale reparameterization for post-training quantization of vision transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17181–17190. IEEE, . 2, 3

[17] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. FQ-ViT: Post-training quantization for fully quantized vision transformer. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1173–1179. International Joint Conferences on Artificial Intelligence Organization. 2

[18] Vien V. Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. 3

[19] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, pages 15682–15694. Curran Associates, Inc., 2021. 2

[20] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization, 2021. 2

[21] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1310–1318. PMLR. ISSN: 1938-7228. 3

[22] Francesco Pinto, Philip Torr, and Puneet K. Dokania. Are vision transformers always more robust than convolutional neural networks? 2

[23] 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. 3

[24] Minho Sim, Jongwhoa Lee, and Ho-Jin Choi. Attention masking for improved near out-of-distribution image detection. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 195–202. ISSN: 2375-9356. 2

[25] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. 2, 3

[26] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357. PMLR, . ISSN: 2640-3498. 2, 3, 4

[27] Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT, . 2, 3

[28] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The INaturalist species classification and detection dataset. pages 8769–8778. 3

[29] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. pages 4921–4930. 3

[30] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. ISSN: 1063-6919. 3

[31] Srikar Yellapragada, Kowshik Thopalli, Vivek Narayanaswamy, Wesam Sakla, Yang Liu, Yamen Mubarka, Dimitris Samaras, and Jayaraman J. Thiagarajan. Leveraging registers in vision transformers for robust adaptation. 4

[32] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. PTQ4vit: Post-training quantization for vision transformers with twin uniform quantization. In *Computer Vision – ECCV 2022*, pages 191–207. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science. 2

[33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. 40(6):1452–1464. 3