# Output Sensitivity-Aware DETR Quantization

**Anonymous submission**


## Abstract

DEtection TRansformer (DETR), the first attempt to utilize transformer in an end-to-end object detection pipeline, achieves promising performance. However, the large model size hinders its deployment to practical applications such as autonomous driving. In this work, we explore a method for post-training mixed-precision quantization of a pretrained DETR model. Specifically, we find previously proposed loss-based sensitivity analysis unsuitable for the DETR model due to the natural roughness of its loss caused by the Hungarian matching process. Thus, we propose a novel distillation-based output sensitivity analysis framework, and evaluate layer sensitivity of an uniformly-quantized DETR model. The proposed technique leads to a mean average precision (mAP) of 40.9%, 36.4%, and 25.9% for, correspondingly, 6-bit, 5-bit and 4-bit quantized DETR models on the MS-COCO dataset. This outperforms the conventional uniform quantization method by 1.1%, 1.3% and 5.2%, respectively.


## Introduction and related work

Object detection is a challenging practical application in computer vision, which is crucial for tasks such as autonomous driving, intelligent surveillance etc. Over the years, multiple major neural network architectures for object detection (Girshick 2015; Ren et al. 2015; Redmon et al. 2016; He et al. 2017) have been discovered.

Recent DEtection TRansformer (DETR) (Carion et al. 2020) with a transformer-based architecture introduces the first model with end-to-end object detection pipeline. This results in a state-of-the-art performance on multiple datasets. However, such performance comes at the cost of large model size and complicated architecture. Therefore, a compression is needed to reduce the memory footprint and inference latency time on mobile and edge devices.

Quantization serves as a popular and effective compression method for numerous object detection models. This architecture-agnostic technique can bring direct memory and computation cost savings on edge devices (Horowitz 2014). Given the different sensitivity of each layer to the weight quantization noise, it is intuitive to apply mixed-precision quantization scheme. Then, a higher precision can be applied to more sensitive layers, while lower bit-width can be reserved for the less sensitive ones. A series of previous works (Dong et al. 2019, 2020; Yang et al. 2021) estimate layer's sensitivity using a Hessian characteristic computed from model's training loss. The exact precision of each layer can then be found using an Integer Linear Programming (ILP) by minimizing overall sensitivity under the model size constraint. This line of work achieves tremendous success in quantizing image classification models, but no attempt has been made to quantize DETR-type detectors.

In this work, we extend conventional loss-based sensitivity analysis to DETR quantization, and make a surprising discovery that it is not an out-of-the-box solution. Unlike classification models with a single output-label mapping for each input image, DETR outputs 100 groups of potential object labels and bounding boxes for each input. Then, they have to be matched to the ground truth boxes before the loss computation (Carion et al. 2020). This matching process introduces a discontinuity in the loss surface with respect to the change of model output, which may hinder the precision of a local Hessian approximation.

To tackle this challenge, we propose an alternative approach to perform Hessian sensitivity analysis, which is based on the difference between the quantized and full-precision model outputs. This approach avoids the matching issue and can be smoothly defined across the entire output space. Previous quantization methods also consider the output difference in the format of distillation, which has been shown to improve the convergence of quantization-aware training (Polino, Pascanu, and Alistarh 2018; Yao et al. 2021). Meanwhile, our work is the first to apply the output difference as a metric for sensitivity analysis in the post-training quantization (PTQ) setting. Furthermore, to better approximate the loss difference induced by quantization, we employ sensitivity analysis on an 8-bit uniformly quantized model instead of the full-precision model. This helps weight values to be closer to ones in the final quantization scheme. We empirically show that the proposed techniques significantly improve the PTQ performance of the DETR model.

## Method

### Output sensitivity-aware quantization scheme

The main challenge in mixed-precision PTQ is to determine the exact precision for each layer in the model. Previous work considers the goal of minimizing the training loss difference between quantized and full-precision models (Dong et al. 2020; Yao et al. 2021). Formally, suppose $N$ predic-

tions are made by the DETR for each image, where each prediction contains the class logits $\hat{p}_i$ and bounding box locations $\hat{b}_i$. The quantization objective can be formulated as

$$\min_{Q} \sum_{i=1}^{N} \mathcal{L}(\hat{p}_{Q_i}, \hat{b}_{Q_i}; p_j, b_j) - \sum_{i=1}^{N} \mathcal{L}(\hat{p}_i, \hat{b}_i; p_j, b_j), \quad (1)$$

where $Q$ denotes the quantization scheme, $\mathcal{L}$ is the training loss function, and $p_j, b_j$ are the $j$-th ground truth class label and bounding box matched to the $i$-th output.

Equation (1) is a discrete nonlinear objective, which makes it infeasible to be efficiently solved directly. Therefore, we express this optimization objective using the Taylor expansion approximation proposed by Dong et al. (2020) as

$$\min_{Q} \sum_{i=1}^{L} \bar{Tr}(H_{W_i}) ||Q(W_i) - W_i||_2^2, \quad (2)$$

where $L$ denotes the number of layers in the model, $W_i$ is the $i$-th layer's weight, and $\bar{Tr}(H_{W_i})$ is the averaged Hessian trace of the DETR training loss $\mathcal{L}$ with respect to the layer weight $W_i$. The Hessian trace can be numerically estimated using the Hutchinson algorithm (Dong et al. 2020) and is well-supported by PyTorch. This approximation disentangles the selection of quantization scheme $Q$ from the nonlinear loss function $\mathcal{L}$, which allows to efficiently solve the optimization objective as an ILP problem.

While this framework is successfully used in quantization of classification models, it underperforms for DETR detector. Note that the formulation of the loss $\mathcal{L}$ of DETR depends on the mapping between model prediction and ground truth boxes, which is a discrete search process (Carion et al. 2020). This can result in a sudden change in the loss surface when the prediction of quantized and full-precision models are mapped to different ground truth bounding boxes. Such a phenomena cannot be captured by the local Hessian information computed using training loss of the full-precision model. Hence, this leads to imprecise sensitivity measurement and improper quantization scheme.

To mitigate the matching process in loss computation, we propose to directly use the difference between quantized model output and full-precision model output, where an one-to-one mapping is available. Specifically, we modify the objective in Equation (1) as

$$\min_{Q} \sum_{i=1}^{N} \mathcal{L}_{dis}(\hat{p}_{Q_i}, \hat{b}_{Q_i}; \hat{p}_i, \hat{b}_i), \quad (3)$$

where $\mathcal{L}_{dis}$ is the distillation loss between the quantized student model and the full-precision pretrained teacher model output, which we define in the next section.

Similar to the approximation in Equation (2), we reformulate the objective in Equation (3) as an ILP problem by computing the Hessian trace of the distillation loss $\mathcal{L}_{dis}$ with respect to the student model weights, where the student model is also floating-point[1].

[1]Distillation loss achieves global minima for two identical floating-point student and teacher models, where both the loss value and loss gradient with respect to student model weights are 0, but Hessian is non-zero.

Furthermore, the approximation in Equation (2) can be more precise when the quantization error $||Q(W_i) - W_i||_2^2$ is small. It is even more important for DETR, where the model architecture and the loss surface is complicated. Therefore, when we search for a quantization scheme with lower precision, we propose to compute the Hessian trace and weight quantization error with respect to the 8-bit quantized weight $W_{qi}$ instead of the full-precision $W_i$. Such technique better models the true loss surface of the final quantized model. Therefore, our final output sensitivity-aware quantization objective can be written by

$$\min_{Q} \sum_{i=1}^{L} \bar{Tr}(H^{dis}_{W_{qi}}) ||Q(W_i) - W_{qi}||_2^2, \quad (4)$$

where $H^{dis}_{W_{qi}}$ is the trace of the output distillation loss computed with respect to a student 8-bit quantized DETR model.

## Multi-layer distillation objective

We carefully design a distillation loss function to measure the output difference for our output sensitivity-aware quantization scheme. We expect the quantized DETR model to have the same behavior as the full-precision pretrained model. Hence, we set our loss such that each pair of $\{\hat{p}_i, \hat{b}_i\}$ from the full model and $\{\hat{p}_{Q_i}, \hat{b}_{Q_i}\}$ from the quantized model to match exactly according to the query index $i$ without performing the Hungarian matching process as in the original DETR loss. We distill the class output using the temperature-scaled KL divergence loss to improve its smoothness (Hinton et al. 2015), and distill the bounding box output with the $\ell_1$ loss as in DETR. Then, our distillation loss can be expressed as

$$\mathcal{L}_{dis} = \lambda_{KL} \mathcal{D}_{KL} \left( \Psi(\hat{p}_{Q_i}/T) || \Psi(\hat{p}_i/T) \right) + ||\hat{b}_{Q_i} - \hat{b}_i||_1, \quad (5)$$

where $\Psi(\cdot)$ denotes the softmax function, $T$ is the distillation temperature, and the hyperparameter $\lambda_{KL}$ balances loss terms. We select $T = 6$ and $\lambda_{KL} = 0.05$ in our experiments from a heuristic search. We use a full-precision pretrained DETR as the teacher for distillation.

Furthermore, the block-wise architecture of the DETR decoder enables each decoder block to generate auxiliary outputs based on its output tokens. Loss functions computed on the auxiliary outputs during training have been found to be helpful in transformer-based architectures (Al-Rfou et al. 2019; Carion et al. 2020). Therefore, we propose to perform a multi-layer distillation where we apply the distillation loss from Equation (5) on the auxiliary outputs of all decoder blocks for quantized and full-precision models. We follow the common practice in DETR, where the feed-forward networks (FFNs) share parameters during computation of auxiliary outputs for all decoder blocks. We accumulate distillation losses from all decoder blocks and use the resulted distillation loss in our sensitivity-aware quantization optimization introduced in Equation (4).

## DETR quantization details

As this work mainly explores the effect of sensitivity analysis objectives on weight quantization, we apply no additional

tricks on the quantizer design in our experiments. We use a symmetric linear quantizer to quantize weight tensor $W$ to $n$ bits, which is defined by

$$Q(W) = \text{Round}\left[W\frac{2^{n-1}-1}{\max(|W|)}\right]\frac{\max(|W|)}{2^{n-1}-1}. \quad (6)$$

We quantize all trainable weights in the DETR model with an exception of the final FFN layers for the class and bounding box outputs. Quantization of these FFN layers leads to catastrophic performance drop in the PTQ setting. We directly apply our method to the official checkpoint of DETR with ResNet-50 backbone[2] without any further training or quantization-aware finetuning. Only weights are quantized while activations are left as full-precision in our experiments. To determine the exact precision of each layer, we use the same ILP solver as in HAWQ-V3 (Yao et al. 2021), which minimizes total layer sensitivity under the constraints of average weight precision of all weight elements. We allow the layer precision to take any value in the range of 3 to 6 when targeting 4-bit and 5-bit quantization, and 4 to 7 for 6-bit quantization.

## Experiments

We evaluate mixed-precision quantization schemes generated with the following three sensitivity metrics:

- **Loss Sensitivity** with Hessian of the training loss;
- **Output Sensitivity Float** with Hessian of the distillation loss with respect to the floating-point weights; and
- **Output Sensitivity Quant** with Hessian of the distillation loss with respect to the uniform 8-bit weights.

All the Hessian traces are computed on the same set of 100 images sampled from the MS COCO 2017 training dataset (Lin et al. 2014). The images are sampled such that the class distribution of objects in the images roughly approximates that of the entire training set. We randomly sampled 10 sets of 100 training images and found that the variance of each set's loss to be less than 1% of the mean. This justifies our sampling strategy for stable loss and Hessian computation.

### Sensitivity analysis and precision assignment

Figure 1 visualizes the log of Hessian trace for each DETR layer estimated using the above three metrics. Figure 2 shows corresponding precision assignment with a target 5-bit average precision found by the ILP solver. We omit other quantization scheme visualizations due to space limitation, but the observed trend is similar. According to Figures 1-2, our output sensitivity method has a more stable Hessian trace estimation, which leads to less sudden drop of precision especially in the backbone layers (the last third part in the figures). Meanwhile, output sensitivity measured using the quantized student model leads to a more uniform trace distribution across the layers and smaller trace values. This demonstrates that the quantized student leads to a better weight approximation and smoother loss surface. The exact reason behind this phenomena is worth further exploration.

[2]https://dl.fbaipublicfiles.com/detr/detr-r50-e632da11.pth

Table 1: PTQ performance of DETR quantized with different schemes on the COCO 2017 dataset. "Precision" denotes the average precision of all trainable weights. $\Delta$ indicates the performance improvements over uniform quantization.

| Precision | Quant scheme | Metric | |
|---|---|---|---|
| | | mAP(%) | $\Delta$ |
| Float | N/A | 42.0 | - |
| 8-bit | Uniform | 41.1 | - |
| 6-bit | Uniform | 39.8 | - |
| | Loss Sen | 36.7 | -3.1 |
| | Out Sen Float | 37.8 | -2.0 |
| | **Out Sen Quant** | **40.9** | **+1.1** |
| 5-bit | Uniform | 35.1 | - |
| | Loss Sen | 26.6 | -8.5 |
| | Out Sen Float | 27.5 | -7.6 |
| | **Out Sen Quant** | **36.4** | **+1.3** |
| 4-bit | Uniform | 20.7 | - |
| | Loss Sen | 19.9 | -0.8 |
| | Out Sen Float | 20.7 | 0.0 |
| | **Out Sen Quant** | **25.9** | **+5.2** |

### PTQ performance

We calculate PTQ mean average precision (mAP) for three mixed-precision quantization schemes using the corresponding sensitivity metric on the COCO validation set in Table 1. We find that unlike classification models, where mixed-precision scheme with loss sensitivity always outperforms uniform quantization, it is not the case for DETR detector due to its loss surface and the resulted rough sensitivity measurements. At the same time, the proposed output-difference sensitivity metric consistently improves mAP. Lastly, our output sensitivity metric estimated using the quantized student leads to even higher performance and outperforms the uniform quantization scheme by a large margin. For comparison, our 6-bit mixed-precision quantization scheme achieves approximately the same mAP as 8-bit uniform scheme, which shows the effectiveness of our approach.

## Conclusions and future work

In this work, we empirically demonstrate that Hessian sensitivity analysis based on the training loss is not suitable for designing quantization schemes of DETR-type models. We show that a multi-layer distillation loss, which measures the output difference between quantized and full-precision models, is a better optimization proxy for mixed-precision sensitivity analysis. Finally, we further improve the PTQ performance using the proposed metric and a uniformly-quantized model as a student instead of a full-precision model.

In future, we will further explore if our findings can be extended to more advanced transformer-based object detection architectures and other datasets. We also plan to find theoretical reasons why the distillation loss computed with respect to a quantized model significantly benefits the sensitivity analysis for DETR model, and if this benefit still holds during the quantization-aware training process. We hope our findings can inspire future research on designing more efficient object detectors for practical real-world applications.

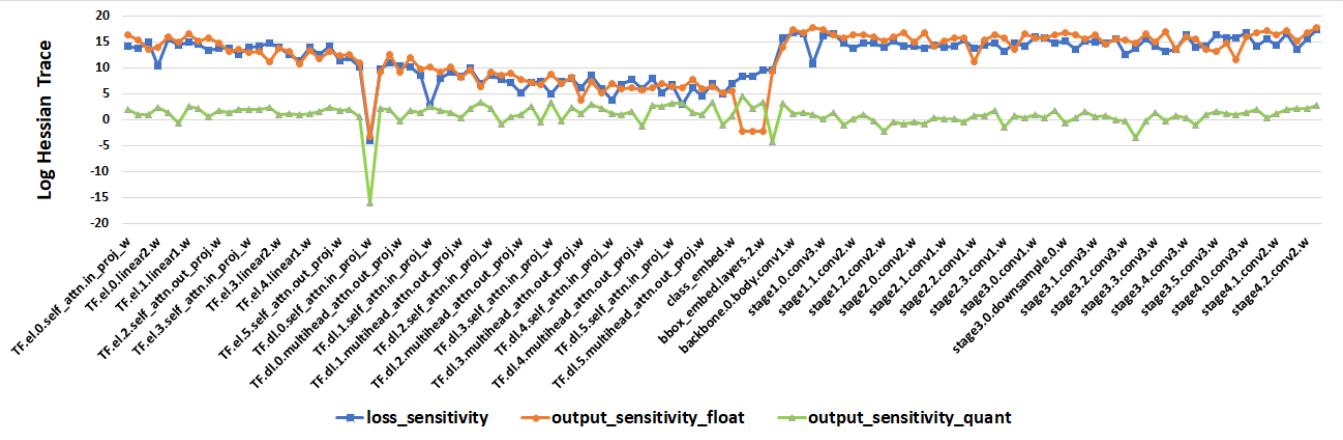


Figure 1: Layer-wise log Hessian trace under different sensitivity metrics.

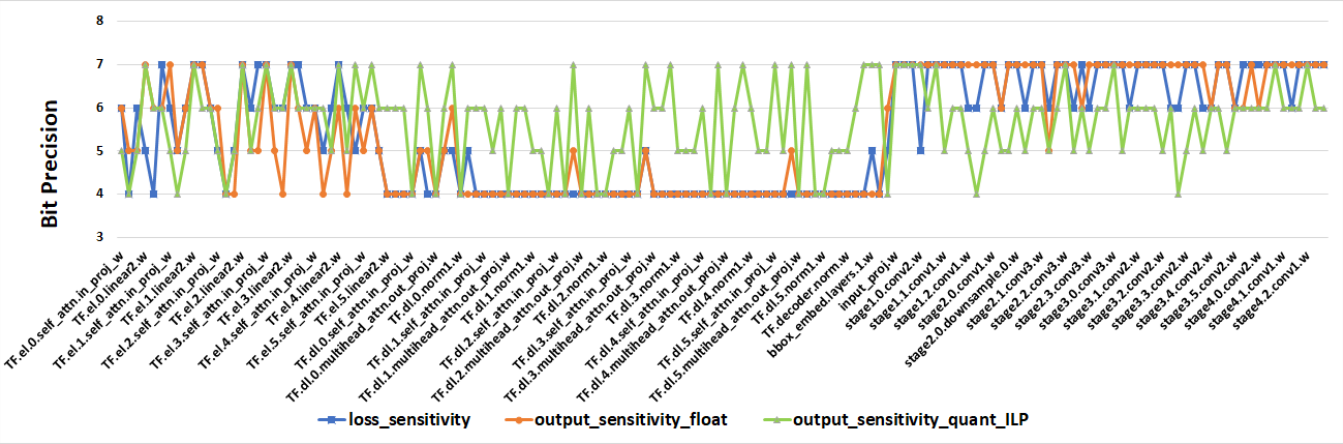


Figure 2: Precision assignments based on the Hessian trace for average 5-bit quantization.

## References


Al-Rfou, R.; Choe, D.; Constant, N.; Guo, M.; and Jones, L. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference*, volume 33, 3159–3166.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229. Springer.

Dong, Z.; Yao, Z.; Arfeen, D.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2020. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33: 18518–18529.

Dong, Z.; Yao, Z.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2019. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE International Conference on Computer Vision*, 293–302.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Horowitz, M. 2014. 1.1 computing's energy problem (and what we can do about it). In *ISSCC*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Polino, A.; Pascanu, R.; and Alistarh, D. 2018. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on CVPR*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*.

Yang, H.; Duan, L.; Chen, Y.; and Li, H. 2021. BSQ: Exploring bit-level sparsity for mixed-precision neural network quantization. *arXiv preprint arXiv:2102.10462*.

Yao, Z.; Dong, Z.; Zheng, Z.; Gholami, A.; Yu, J.; Tan, E.; Wang, L.; Huang, Q.; Wang, Y.; Mahoney, M.; et al. 2021. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, 11875–11886. PMLR.