InfoQ: Mixed-Precision Quantization via Global Information Flow

Mehmet Emre Akbulut, Hazem Hesham Yousef Shalby, Fabrizio Pittorino, and Manuel Roveri

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
mehmetemre.akbulut@mail.polimi.it, {hazemhesham.shalby, fabrizio.pittorino, manuel.roveri}@polimi.it

Abstract

Mixed-precision quantization (MPQ) is crucial for deploying deep neural networks on resource-constrained devices, but finding the optimal bit-width for each layer represents a complex combinatorial optimization problem. Current state-of-the-art methods rely on computationally expensive search algorithms or local sensitivity heuristic proxies like the Hessian, which fail to capture the cascading global effects of quantization error. In this work, we argue that the quantization sensitivity of a layer should not be measured by its local properties, but by its impact on the information flow throughout the entire network. We introduce InfoQ, a novel framework for mixed-precision quantization that is training-free in the bit-width search phase. InfoQ assesses layer importance by performing a single forward pass to measure the change in mutual information in the remaining part of the network, thus creating a global sensitivity score. This approach directly quantifies how quantizing one layer degrades the information characteristics of subsequent layers. The resulting scores are used to formulate bit-width allocation as an integer linear programming problem, which is solved efficiently to minimize total sensitivity under a given budget (e.g., model size or BitOps). Our retraining-free search phase provides a superior search-time/accuracy trade-off (using two orders of magnitude less data compared to state-of-the-art methods such as LIMPQ), while yielding up to a 1% accuracy improvement for MobileNetV2 and ResNet18 on ImageNet at high compression rates (14.00× and 10.66×).

Code — https://github.com/mehmetemreakbulut/InfoQ
Extended version — https://arxiv.org/abs/2508.04753

1 Introduction

The significant computational demands of Deep Neural Networks (DNNs) are a primary barrier to their deployment on resource-constrained edge devices. Mixed-precision quantization (MPQ) is a key technique to overcome this, assigning low bit-widths to robust layers while preserving precision for sensitive ones to outperform uniform quantization (Dong et al. 2023). However, identifying the optimal bit-width configuration for a network is a combinatorial problem; the search space grows exponentially with network depth, making an exhaustive search infeasible.

Current approaches to this challenge fall into two categories, both of which suffer from fundamental limitations. *Search-based* methods (Wang et al. 2019; Wu et al. 2018) use expensive reinforcement learning or neural architecture search (NAS) to find good configurations. While sometimes effective, they are computationally prohibitive and treat the network as a black box, failing to build a principled model of quantization impact. More efficiently, *criterion-based* methods use a proxy metric to score layer sensitivity in a single shot. A relevant class of these methods use the trace of the Hessian (Dong et al. 2019b,a). The intuition is that layers in *flat* regions of the loss landscape are less sensitive. However, the prohibitive cost of the full Hessian forces these methods to use a *block-diagonal approximation*, which explicitly ignores all inter-layer dependencies. This makes the Hessian-based analysis fundamentally local, failing to capture how quantization error in one layer cascades and degrades the representational quality of subsequent layers. Other criteria, like learned step-sizes (Tang et al. 2023), are similarly layer-local and often require expensive retraining. This leaves a critical gap: a fast, principled MPQ method that evaluates sensitivity based on its global impact across layers.

In this work, we argue that the true sensitivity of a layer is not an intrinsic, local property, but is defined by the *global impact* it has on the *information flow throughout the entire network*. We introduce *InfoQ*, a framework that directly builds upon this principle. Instead of using local proxies, we ask a more fundamental question: *How does quantizing a layer degrade the ability of downstream layers to represent critical information about the input and the final task?*

To answer this, *InfoQ* leverages tools from information theory. For each layer and candidate bit-width, we perform a single forward pass and measure the resulting change in mutual information (MI) at subsequent *observer* layers. This change directly quantifies the global, cascading effect of the quantization error. These global sensitivity scores are then used to formulate the bit-width allocation as an Integer Linear Programming (ILP) problem, which is solved efficiently to find the optimal configuration under a given resource budget (e.g., model size or BitOps). Our search phase is entirely training-free and provides a global, interpretable sensitivity measure, in contrast to existing local methods. Our contributions are:

1. We introduce *InfoQ*, the first MPQ framework to use a

Method	AutoQ	DNAS	LIMPQ	HAWQv2	MPQCO	*InfoQ* (Ours)
Iterative Search Avoidance	✗	✗	✓	✓	✓	✓
Unlimited Search Space	✓	✗	✓	✓	✗	✓
Training-Free Search	✗	✗	✗	✓	✓	✓
Fully Automatic Bit Assignment	✓	✓	✓	✓	✗	✓
Global Impact Metric	✗	✗	✗	✗	✗	✓

Table 1: Comparison of key properties in state-of-the-art MPQ methods. *InfoQ* is the only method that combines a training-free search with a global sensitivity metric, enabling fully automatic bit allocation that accounts for network-wide error propagation.

global sensitivity metric based on information flow, directly capturing the cascading effects of quantization error ignored by local methods.

2. Our method achieves state-of-the-art accuracy, recovering full-precision accuracy on ResNet18 and narrowing the gap to just 0.34% on ResNet50 at high compression rates, all while using orders of magnitude less data for the search phase than competing methods at the same accuracy level.
3. Using a practical method for mutual information estimation, we analyze deep networks compression from an information-theoretic perspective, highlighting a valuable application for both the model compression and information theory literature.

2 Related Work

Mixed-precision quantization (MPQ) consistently demonstrates superior performance over uniform-precision approaches (Dong et al. 2023). The central challenge in MPQ is navigating the combinatorial search space of bit-width configurations. Methodologies to solve this problem are broadly categorized as *search-based* and *criterion-based*.

The *search-based* methods solve MPQ problem with computationally intensive search algorithms. Reinforcement learning (RL) has been employed to select layer-wise (Wang et al. 2019) or even kernel-wise (Lou et al. 2020) bit allocations. Another explored direction adapts techniques from Neural Architecture Search (NAS), such as differentiable search over a super-network of quantization operators (Wu et al. 2018; Cai and Vasconcelos 2020; Yu et al. 2020). While these methods can yield high-performing configurations, their reliance on costly, iterative search (often requiring a significant number of GPU-hours) and large validation datasets makes them impractical for many applications. Our work bypasses this expensive search phase entirely.

To circumvent the cost of search, *criterion-based* methods use an efficient proxy metric for layer sensitivity to guide a one-shot bit allocation. Methods like HAWQ (Dong et al. 2019b) and HAWQv2 (Dong et al. 2019a) use the trace or eigenvalues of the Hessian as a proxy for layer sensitivity, using second-order information from the loss landscape. As noted in the Introduction, these methods are constrained to a block-diagonal approximation of the Hessian, making their analysis fundamentally local. Other works have proposed alternative proxies, including layer-wise orthogonality metrics (Ma et al. 2021) and the learned scale parameters of quantizers (Tang et al. 2023). While significantly faster than search-based methods, these approaches share a critical limitation: their reliance on layer-local heuristics prevents them from quantifying the global, cascading effects of quantization error. *InfoQ* is a criterion-based method, but its criterion is, by design, global.

Concerning the grounding of model compression in information theory, it is a promising direction for developing more principled methods. The Information Bottleneck (IB) principle, in particular, offers a theoretical framework for analyzing the trade-off between compression and predictive accuracy (Tishby and Zaslavsky 2015). The IB framework has been successfully applied to guide network pruning, where channels or weights are removed based on their information content (Guo et al. 2023; Zheng et al. 2021; Nielsen et al. 2021). However, its application to MPQ has remained limited.

Our work, *InfoQ*, is a criterion-based method that, for the first time, successfully uses information-theoretic tools to construct a *practical global sensitivity criterion* for MPQ. Unlike prior IB-based model pruning methods, we do not attempt a complex optimization of the IB objective. Instead, we use mutual information as a direct, interpretable metric to measure the end-to-end impact of quantization. As summarized in Table 1, *InfoQ* is the first MPQ framework able to provide a fully automatic bit allocation based on a training-free, interpretable, and global analysis of the network information flow.

3 Background: Estimating Mutual Information in Deep Networks

To analyze the global impact of quantization, we turn to concepts from information theory, particularly those used by the Information Bottleneck (IB) principle (Tishby and Zaslavsky 2015). The IB principle models a DNN as a system that learns compressed representations of an input X. The quality of a representation L_i at layer i is evaluated by two fundamental quantities: the mutual information it retains about the input, $I(X; L_i)$, and the information it preserves about the final target, $I(L_i; Y)$ (Geiger 2020). While the IB principle aims to optimize the trade-off between these two, we leverage them as direct probes to measure the information variations caused by quantization along the network. For two continuous random variables U and V, the mutual information $I(U; V)$ quantifies their statistical dependency and is defined as $I(U; V) = \int_{\mathcal{V}} \int_{\mathcal{U}} p(u, v) \log \frac{p(u,v)}{p(u)p(v)} du dv$, where $p(u, v)$ is the joint probability density function and $p(u)$ and $p(v)$ are the marginal densities. The quantities

$I(X; L_i)$ and $I(L_i; Y)$ are the core metrics for quantifying the informational properties of a network representations.

There are however fundamental challenges in estimating mutual information for high-dimensional, deterministic functions like DNNs. The first obstacle is that the activations L_i are a deterministic function of the input X. In a continuous setting, this leads to a theoretically infinite mutual information, $I(X; L_i)$, making the metric uninformative (Amjad and Geiger 2018). The second is that, even when non-determinism is introduced (e.g., through quantization (Lorenzen, Igel, and Nielsen 2021) or noise injection), the high dimensionality of both the input space $\mathcal{X}$ and the activation space $\mathcal{L}_i$ makes non-parametric MI estimation intractable (the *curse of dimensionality*): the number of samples required for a reliable estimate grows exponentially with data dimension (Goldfeld et al. 2019).

To overcome these challenges, our work leverages two key techniques from recent literature on tractable MI estimation. The first is *lossy compression* of the input space (Butakov et al. 2023), where a high-dimensional input X (e.g., an image) is mapped to a lower-dimensional, semantically rich feature vector using a powerful pre-trained encoder. This reduces the dimensionality of one of the variables, mitigating the estimation problem. The second central technique is *Sliced Mutual Information* (SMI), a computationally efficient and scalable surrogate for MI (Goldfeld et al. 2022). Instead of estimating the MI between two high-dimensional random vectors directly, SMI computes the average MI between one-dimensional random projections of these vectors. For vectors $U \in \mathbb{R}^d$ and $V \in \mathbb{R}^p$, SMI is defined as $\text{SMI}(U; V) = \mathbb{E}_{\theta_u, \theta_v}\left[\text{MI}(\theta_u^T U; \theta_v^T V)\right]$, where the projection directions θ_u and θ_v are drawn randomly from unit spheres. The MI between these one-dimensional scalar projections can be robustly and efficiently estimated using standard non-parametric methods, such as Kraskov–Stögbauer–Grassberger (KSG) estimator (Kraskov, Stögbauer, and Grassberger 2003).

4 Methodology

Problem Formalization

Let a deep neural network be defined by a sequence of L layers with full-precision parameters $\boldsymbol{W}$. For mixed-precision quantization, we define a discrete set of candidate bit-widths $\boldsymbol{B} = \{b_0, b_1, \ldots, b_{n-1}\}$. For each layer $\ell \in \{1, \ldots, L\}$, we assign a bit-width for its weights, $b_w^{(\ell)} \in \boldsymbol{B}$, and its activations, $b_a^{(\ell)} \in \boldsymbol{B}$. A complete bit-width configuration for the network is an assignment vector $\boldsymbol{s} = \{(b_w^{(1)}, b_a^{(1)}), \ldots, (b_w^{(L)}, b_a^{(L)})\}$. The set of all possible configurations, $\mathcal{A}$, forms the search space for the MPQ problem. For a typical network with $L = 50$ and $|\boldsymbol{B}| = 4$ candidate bit-widths for both weights and activations, the size of this search space is $|\mathcal{A}| = (|\boldsymbol{B}|^2)^L = 16^{50}$.

The objective of MPQ is to find an optimal bit-width configuration $\boldsymbol{s}^* \in \mathcal{A}$ that minimizes a task-specific loss $\mathcal{L}$ subject to one or more resource constraints:

$$\begin{aligned} \boldsymbol{s}^* = \arg\min_{\boldsymbol{s} \in \mathcal{A}} \quad & \mathcal{L}(f(\boldsymbol{x}; \boldsymbol{s}, \boldsymbol{W_s}), \boldsymbol{y}) \\ \text{subject to} \quad & \text{Cost}(\boldsymbol{s}) \leq C \end{aligned} \tag{1}$$

where $f(\cdot)$ is the network model with parameters $\boldsymbol{W_s}$ and activations quantized according to configuration $\boldsymbol{s}$, $(\boldsymbol{x}, \boldsymbol{y})$ are the data and labels, and $\text{Cost}(\boldsymbol{s})$ represents a resource budget. Common cost functions include model size (Uhlich et al. 2019) or computational complexity, measured in BitOps (Yang and Jin 2020):

$$\text{Cost}(\boldsymbol{s}) = \sum_{\ell=1}^{L} \text{Size}(b_w^{(\ell)}) \quad \text{or} \quad \sum_{\ell=1}^{L} \text{BitOps}(b_w^{(\ell)}, b_a^{(\ell)}) \tag{2}$$

The *InfoQ* Method

Directly solving the optimization problem in Eq. 1 is intractable: evaluating the loss for each candidate $\boldsymbol{s}$ would require retraining or fine-tuning the model, and the combinatorial size of $\mathcal{A}$ makes exhaustive search impossible. We therefore propose a proxy-based method to efficiently estimate the impact of a given bit-width configuration without full model retraining, *InfoQ*, which reformulates the problem as a three-step process. First, it defines a layer quantization sensitivity as the global information degradation it causes throughout the network. Then, it computes this sensitivity for each layer and candidate bit-width using an efficient SMI-based algorithm. Finally, it uses these scores to formulate the bit-width allocation as an Integer Linear Programming (ILP) problem, which can be solved efficiently for any given resource constraint.

Defining Sensitivity via Global Information Flow Our central hypothesis is that the impact of quantizing a layer is not a local phenomenon but is best measured by its effect on the information propagated through subsequent layers. We quantify this impact by measuring the change in the two fundamental information-theoretic quantities analyzed in the IB context: the *input information* $I(X; L_j)$, which measures how much information the representation L_j at a downstream layer j retains about the original input X; and the *task-relevant information* $I(L_j; Y)$, which measures how much information L_j contains about the final task labels Y. A significant change in either of these quantities at a downstream layer j after quantizing an upstream layer i indicates that layer i is sensitive to quantization. To make the estimation of these quantities tractable for high-dimensional data, we employ the SMI surrogate, as introduced in Section 3. Furthermore, we apply *lossy compression* where input images X are mapped to a lower-dimensional embedding $X_{\mathcal{E}}$ using a pre-trained DINOv2 encoder (Oquab et al. 2023) that we denote by $\mathcal{E}$, i.e. $X_{\mathcal{E}} = \mathcal{E}(X)$.

Our sensitivity metric is therefore based on the measured informational degradation in $\text{SMI}(X_{\mathcal{E}}; L_j)$ and $\text{SMI}(L_j; Y)$ relative to a high-precision baseline model (all layers at 8-bit). Let $L_{j,\text{8bit}}$ be the activations of a downstream *observer* layer j in the baseline model, and let $L_{j,b}$ be the activations at the same layer when an upstream layer i has been quantized to bit-width b. We formally define the absolute change in SMI as:

$$\begin{aligned} \Delta\text{SMI}_{X,L}^{(i,b,j)} &= |\text{SMI}(X_{\mathcal{E}}; L_{j,\text{8bit}}) - \text{SMI}(X_{\mathcal{E}}; L_{j,b})| \\ \Delta\text{SMI}_{L,Y}^{(i,b,j)} &= |\text{SMI}(L_{j,\text{8bit}}; Y) - \text{SMI}(L_{j,b}; Y)| \end{aligned} \tag{3}$$

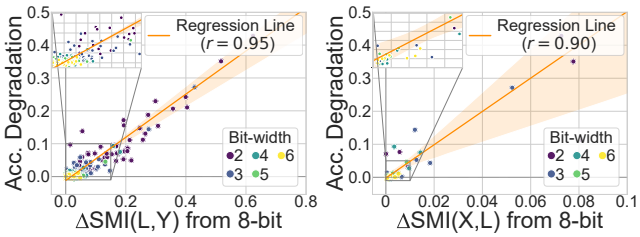

Figure 1: Correlation analysis for ΔSMI metrics defined in Eqs. 3 and accuracy degradation on ResNet50 on ImageNet. The accuracy drop is shown in function of (left panel) the target-relevant information change, $\Delta\text{SMI}(L, Y)$, on the *global average pooling* layer, and (right panel) the input-relevant information change, $\Delta\text{SMI}(X, L)$, on the *conv* layer at index 7. The strong positive correlation demonstrates that the ΔSMI metrics are an effective proxy for quantization sensitivity.

The total sensitivity score for quantizing layer i to bit-width b is the normalized sum of these measured changes across a set of pre-selected *observer layers*.

Observer Layer Selection and Validation A critical component of our method is the selection of downstream *observer layers*, i.e. the locations where we measure the the absolute change in SMI in Eqs. 3. An ideal observer layer is one where a change in its informational content is highly predictive of the final degradation in task performance. We identify these observers for a given network architecture via a one-time, empirical correlation analysis. For each layer i in the network, we quantize it to a low bit-width (<8-bit) while keeping all other layers at 8-bit. We then perform a forward pass and record two values: (1) the final top-1 accuracy degradation, and (2) $\Delta\text{SMI}_{X,L}$ and $\Delta\text{SMI}_{L,Y}$ as per Eqs. 3, at all subsequent layers $j > i$. By repeating this for all quantizable layers, we are able to compute the Pearson correlation coefficient between the absolute change in SMI as per Eqs. 3 and the final accuracy drop. This allows us to identify a robust set of observers by selecting layers strongly correlated with the final accuracy drop (Pearson's $r > 0.70$). This analysis, detailed in the Appendix, yields a clear and consistent heuristic. We find that the change in task-relevant information, $\Delta\text{SMI}_{L,Y}$, measured at layers toward the end of the network (e.g., global pooling and fully-connected layers) exhibits the strongest correlation with accuracy degradation. Conversely, the change in input-relevant information, $\Delta\text{SMI}_{X,L}$, is most predictive when measured at intermediate layers.

As shown in Fig. 1, the absolute change in SMI measured at selected observer layers strongly correlates with the final model accuracy, validating their use as an effective proxy for quantization sensitivity. Fig. 2 visualizes the final sensitivity scores computed for each layer across various bit-widths, which form the basis of our bit allocation strategy.

Sensitivity Scoring and Optimal Bit Allocation With the observer layers identified, we proceed to compute a sensitivity score for each quantizable layer and candidate bit-width.

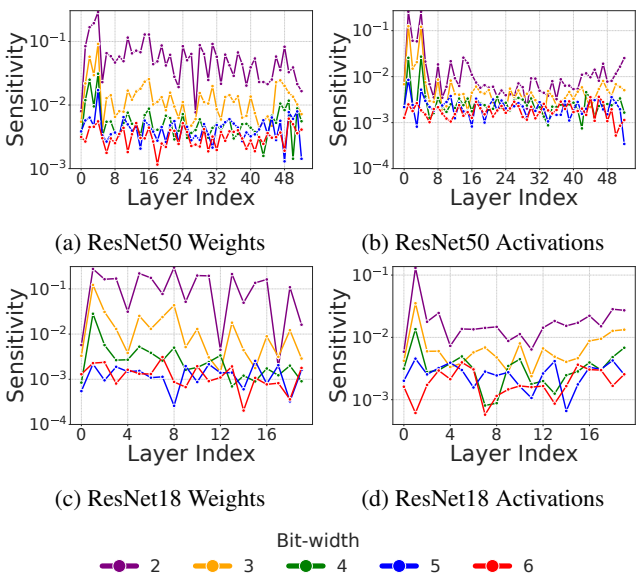

Figure 2: Sensitivity profiles as per Eq. 4 for ResNet50 and ResNet18 on ImageNet in function of the layer index, for different quantization bit-widths.

This process, detailed in Algorithm 1, is designed to be efficient and training-free. The core of the algorithm is a systematic measurement of informational degradation. We first establish a baseline by computing the values $\text{SMI}(X_{\mathcal{E}}; L_j)$ and $\text{SMI}(L_j; Y)$ at all identified observer layers j and k (notice that in general they differ) for a reference model where all layers are quantized to 8-bits. Then, to compute the sensitivity of a target layer i for a candidate bit-width $b \in \boldsymbol{B}$, we perturb this baseline: we quantize only layer i to bit-width b and perform a single forward pass over a small calibration dataset. The sensitivity score is then the normalized sum of the absolute change in SMI (as defined in Eqs. 3) across all downstream observer layers. This procedure is repeated for every (layer, bit-width) pair. Therefore, the final score S for quantizing the parameters of layer i (weights or activations) to bit-width b is formulated as:

$$S(i, b, \mathcal{O}_{XL}^{j>i}, \mathcal{O}_{LY}^{j>i}) = \frac{1}{b} \frac{\langle \Delta\text{SMI} \rangle_{\mathcal{O}^{j>i}}}{\langle \text{SMI} \rangle_{\mathcal{O}^{j>i}}^{8\text{bit}}} \tag{4}$$

$$\langle \Delta\text{SMI} \rangle_{\mathcal{O}^{j>i}} = \sum_{j \in \mathcal{O}_{XL}^{j>i}} \Delta\text{SMI}_{X,L}^{(i,b,j)} + \sum_{k \in \mathcal{O}_{LY}^{j>i}} \Delta\text{SMI}_{L,Y}^{(i,b,k)}$$

$$\langle \text{SMI} \rangle_{\mathcal{O}^{j>i}}^{8\text{bit}} = \sum_{j \in \mathcal{O}_{XL}^{j>i}} \text{SMI}_{X,L}^{(j),8\text{bit}} + \sum_{k \in \mathcal{O}_{LY}^{j>i}} \text{SMI}_{L,Y}^{(k),8\text{bit}}$$

where $\mathcal{O}_{XL}^{j>i}$ and $\mathcal{O}_{LY}^{j>i}$ are the downstream observer layers (with respect to layer i) for $\text{SMI}(X_{\mathcal{E}}; L_j)$ and $\text{SMI}(L_j; Y)$, respectively. The score is normalized by the total baseline information to ensure comparability across different observer sets. We also include a $1/b$ term to penalize lower bit-widths, in order to explicitly account for their inherently higher sensitivity to quantization error. When the ILP solver seeks

to minimize total network sensitivity under a budget, this penalty ensures that, for a comparable ΔSMI value, a higher bit-width configuration is favored, promoting a more stable quantization setup. We study the impact of the factor $1/b$ in the Appendix, showing that the global SMI metric is the dominant factor in our method performance.

Once the sensitivity scores $S_w(\ell, b)$ and $S_a(\ell, b)$ are computed for all layers and bit-widths, we can find the optimal network bit-width configuration $\boldsymbol{s}^*$ by solving a constrained optimization problem. The goal is to find the configuration that minimizes the total network sensitivity while respecting a resource budget C. This is formulated as a classic Integer Linear Programming (ILP) problem:

$$\begin{aligned} \boldsymbol{s}^* = \arg\min_{\boldsymbol{s}\in\mathcal{A}} \quad & \sum_{\ell=1}^{L} \left(S_w(\ell, b_w^{(\ell)}) + \alpha S_a(\ell, b_a^{(\ell)}) \right) \\ \text{subject to} \quad & \text{Cost}(\boldsymbol{s}) \leq C \end{aligned} \tag{5}$$

Here, α is a hyperparameter balancing the relative importance of weight versus activation quantization sensitivity. Cost($\boldsymbol{s}$) is the resource-specific cost function (e.g., model size or BitOps, as in Eq. 2). For only two experiments on ResNet18 and ResNet50, we set $\alpha = 0.1$ and $\alpha = 2$, respectively, to ensure a fair comparison with our main competitor, LIMPQ (Tang et al. 2023). Other experiments on ResNet18 and MobileNetV2 employ weight-only MPQ which does not need hyperparameter α. This standard ILP problem is solved efficiently using an off-the-shelf solver (Mitchell, O'Sullivan, and Dunning 2011), making the final allocation step extremely fast for any given budget C.

Algorithm 1: InfoQ Sensitivity Score Computation

Input: Pre-trained model f, calibration data $(\boldsymbol{X_c}, \boldsymbol{Y_c})$, candidate bit-widths $\boldsymbol{B}$, observer layers sets $\mathcal{O}_{XL}, \mathcal{O}_{LY}$.
Output: Sensitivity scores $S(\ell, b)$ for each layer ℓ and bit-width b.

 # 1. Compute Baseline Information on 8-bit model
Let $\boldsymbol{s}_{\text{8bit}}$ be the configuration with all layers at 8-bit.
Get baseline activations $\{\boldsymbol{L}_{j,\text{8bit}}\}_{j\in\mathcal{O}_{XL}\cup\mathcal{O}_{LY}}$ from $f(\boldsymbol{X_c}; \boldsymbol{s}_{\text{8bit}})$.
$I_{X,L}^{(j,\text{8bit})} \leftarrow \text{SMI}(\boldsymbol{X_c}, \boldsymbol{L}_{j,\text{8bit}})$ for observers $j \in \mathcal{O}_{XL}$
$I_{L,Y}^{(j,\text{8bit})} \leftarrow \text{SMI}(\boldsymbol{L}_{j,\text{8bit}}, \boldsymbol{Y_c})$ for observers $j \in \mathcal{O}_{LY}$
 # 2. Compute Sensitivity for each (layer, bit-width) pair
for each quantizable layer $\ell = 1, \ldots, L$ **do**
  for each bit-width $b \in \boldsymbol{B}$ **do**
    # Perturb model by quantizing only layer ℓ to bit b
    $\boldsymbol{s}_{\text{pert}} \leftarrow \boldsymbol{s}_{\text{8bit}}$
    Set layer ℓ to bit-width b in $\boldsymbol{s}_{\text{pert}}$.
    Get perturbed activations $\{\boldsymbol{L}_{j,b}\}_{j>\ell}$ from $f(\boldsymbol{X_c}; \boldsymbol{s}_{\text{pert}})$.
    # Calculate normalized score as per Eq. 4
    $S(\ell, b, \mathcal{O}_{XL}^{j>i}, \mathcal{O}_{LY}^{j>i}) \leftarrow$ Eq. (4)
  end for
end for
return S

5 Experiments

We conduct a comprehensive evaluation of the *InfoQ* framework across three dimensions. First, we present a state-of-the-art (SOTA) comparison on standard benchmarks after quantization-aware training (QAT). Second, we analyze the post-quantization accuracy *without* QAT to directly assess the quality of our sensitivity metric (Deng et al. 2023). Finally, we evaluate the search efficiency of our method.

State-of-the-Art Comparison with QAT

To evaluate the effectiveness of the bit configurations found by *InfoQ*, we follow standard community practices. We perform experiments on the ImageNet (ILSVRC12) dataset (Deng et al. 2009) using three widely-used architectures: ResNet18/50 (He et al. 2015) and MobileNetV2 (Sandler et al. 2019). For a given model and hardware constraint (e.g., model size or BitOps), we first use *InfoQ* to determine the optimal bit-width configuration. We then apply QAT using the Learned Step-Size (LSQ) method (Esser et al. 2019) to recover accuracy. The candidate bit-width set for both weights and activations is $\boldsymbol{B} = \{2, 3, 4, 5, 6, 7, 8\}$. We compare *InfoQ* against a wide range of leading uniform-precision and mixed-precision methods.

As shown in Table 2, *InfoQ* demonstrates superior performance on ResNet18. In a weight-only quantization setting (avg. 3-bit weights, 8-bit activations), our method achieves an accuracy of 70.94%, surpassing the full-precision baseline and outperforming other methods like OMPQ and MPQCO. In the more challenging setting of joint weight and activation quantization under a 23.04 G-BitOps constraint, *InfoQ* achieves 69.99% top-1 accuracy, marking the smallest accuracy degradation (-0.61%) among all listed methods. This result is particularly strong when compared to LIMPQ, which reports a similar BitOps budget but achieves it by leaving the first layer activations at 32-bit, a significant deviation from a fully quantized model.

On the larger ResNet50 model, under a strict 12.2× weight compression constraint, *InfoQ* achieves a state-of-the-art top-1 accuracy of 77.03% (Table 3). This result represents an accuracy drop of only -0.34% from its full-precision baseline, outperforming prominent methods like HAWQv2 (-1.50%) and LIMPQ (-0.60%). Furthermore, this result is achieved with a training-free search phase that uses only 6000 calibration samples, in contrast to retraining-based proxies like LIMPQ which require a substantial fraction of the full training set.

To evaluate performance on a non-residual, depthwise-separable architecture, we test on MobileNetV2. We set an aggressive 14.00× weight compression target to ensure quantization is applied across the entire network, not just the final fully-connected layer. As shown in Table 4, *InfoQ* excels in this setting, achieving 69.83% accuracy. This significantly outperforms prior methods such as HAQ (-5.12% drop) and MPQCO (-3.36% drop) at a comparable compression rate.

Direct Evaluation of the Sensitivity Metric

The performance of a model after QAT depends on both the quality of the bit-width configuration and the effectiveness

Method	W-bit	A-bit	Top-1/Full	Top-1/Quant	Top-1/Drop	BitOps(G)	Searching Data
LQ-Nets (Zhang et al. 2018)	3	32	70.30	69.30	-1.00	-	-
LQ-Nets (Zhang et al. 2018)	4	32	70.30	70.00	-0.30	-	-
OMPQ (Ma et al. 2021)	3MP	8	71.08	69.94	-1.14	-	**64**
MPQCO (Chen, Wang, and Cheng 2021)	3MP	32	69.76	69.50	-0.26	-	1024
MPQCO (Chen, Wang, and Cheng 2021)	3MP	8	69.76	69.39	-0.37	-	1024
InfoQ (Ours)	**3MP**	**8**	**70.60**	**70.94**	**+0.34**		4800
PACT (Choi et al. 2018)	3	3	70.40	68.10	-2.30	23.09	-
LQ-Nets (Zhang et al. 2018)	3	3	70.30	68.20	-2.10	23.09	-
Nice (Baskin et al. 2021)	3	3	69.80	67.70	-2.10	23.09	-
AutoQ (Lou et al. 2020)	3MP	3MP	69.90	67.50	-2.40	-	-
SPOS (Guo et al. 2020)	3MP	3MP	70.90	69.40	-1.50	21.92	-
DNAS (Wu et al. 2018)	3MP	3MP	71.00	68.70	-2.30	25.38	-
LIMPQ[1](Tang et al. 2023)	3MP	3MP	70.50	69.70	- 0.80	23.07	0.6M
InfoQ (Ours)	**3MP**	**3MP**	**70.60**	**69.99**	**-0.61**	**23.04**	**4800**

[1] First layer activation is not quantized.

Table 2: Results for ResNet18 on ImageNet under BitOps constraints. Upper part is weight-only, below part is non-limited MPQ. 'W-bit' and 'A-bit' represent the bit-widths of weights and activations, respectively. 'MP' denotes mixed-precision quantization. 'Top-1/Quant' and 'Top-1/Full' refer to the top-1 accuracy of the quantized and full-precision models. 'Top-1/Drop' is defined as the accuracy drop: Top-1/Full - Top-1/Quant.

Method	W-bit	A-bit	Top-1/Full	Top-1/Quant	Top-1/Drop	W-C	Searching Data
PACT (Choi et al. 2018)	3	3	76.90	75.30	-1.60	10.67×	-
LQ-Nets (Zhang et al. 2018)	3	3	76.00	74.20	-1.80	10.67×	-
HAQ (Wang et al. 2019)	3MP	8	75.30	76.20	-0.90	10.57×	-
DiffQ (Défossez, Adi, and Synnaeve 2022)	3MP	32	77.10	76.30	-0.80	11.10×	-
BP-NAS (Yu et al. 2020)	4MP	4MP	77.50	76.70	-0.80	**11.10**×	-
HAWQ (Dong et al. 2019b)	MP	MP	77.30	75.50	-1.80	12.20×	-
HAWQv2 (Dong et al. 2019a)	MP	MP	77.30	75.80	-1.50	12.20×	**256**
MPQCO (Chen, Wang, and Cheng 2021)	MP	4MP	76.10	75.30	-0.80	12.20×	1024
LIMPQ[1](Tang et al. 2023)	3MP	4MP	77.50	76.90	-0.60	12.20×	0.6M
InfoQ (Ours)	**3MP**	**4MP**	**77.37**	**77.03**	**-0.34**	12.20×	6000

[1] First layer activation is not quantized.

Table 3: Results for ResNet50 on ImageNet under BitOps constraints. 'W-C' denotes the weight compression rate.

Method	W-bit	A-bit	Top-1/Full	Top-1/Quant	Top-1/Drop	W-C	Searching Data
DC (Han, Mao, and Dally 2015)	MP	32	71.87	58.07	-13.80	13.93×	-
HAQ (Wang et al. 2019)	MP	32	71.87	66.75	-5.12	**14.07**×	-
MPQCO (Chen, Wang, and Cheng 2021)	MP	8	71.88	68.52	-3.36	13.99×	**1024**
InfoQ (Ours)	**MP**	**8**	**72.02**	**69.83**	**-2.19**	14.00×	**1024**

Table 4: Results for MobileNetv2 on ImageNet under weight compression constraints, denoted by 'W-C'.

of the fine-tuning process. To isolate and directly evaluate the quality of the bit allocation produced by our sensitivity metric, we compare its performance *before* any QAT, a setup often referred to as post-training quantization (PTQ) (Deng et al. 2023). A superior PTQ accuracy indicates that the underlying sensitivity metric is more effective at identifying a robust quantization configuration.

We compare *InfoQ* against leading criterion-based methods (HAWQv2, MPQCO, LIMPQ) across a range of model size constraints for ResNet18/34/50 on ImageNet. For each method and each constraint, we generate the optimal bit-width configuration and report the corresponding PTQ accuracy. The results are presented in Figure 4. In high-compression regimes (e.g., model sizes approaching that of uniform 3-bit quantization), the performance gap between methods becomes most apparent. In these scenarios, *InfoQ* consistently outperforms all other methods, often by a significant margin. For instance, on ResNet18 at a 4.5MB constraint, *InfoQ* yields a configuration that is 25% more accurate than its closest competitor. This demonstrates that our global, information-based sensitivity metric is fundamentally more effective at preserving model accuracy than local, Hessian-based or retraining-based proxies. Furthermore, a superior initial configuration provides a better starting point

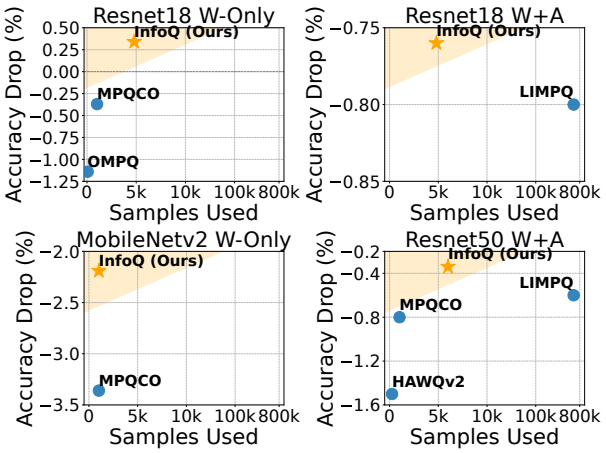

Figure 3: Performance of our method against state-of-the-art *criterion-based* methods. 'W-Only' and 'W+A' mean 'weight-only' and 'weight+activation' MPQ.

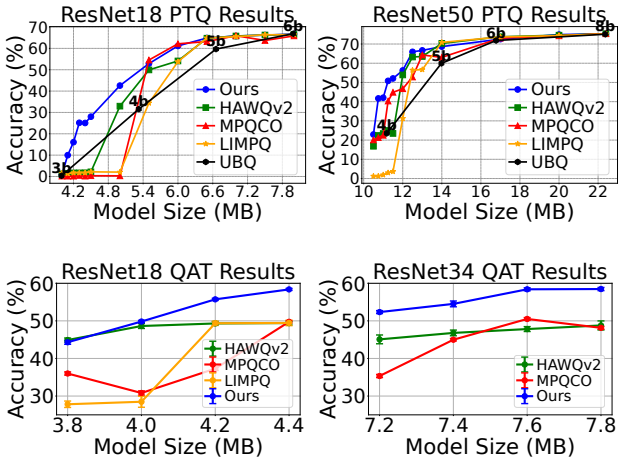

Figure 4: Direct evaluation of our method against state-of-the-art *criterion-based* methods on ResNet18 and ResNet50 (top). The results are derived from best-performing bit-width configurations according to sensitivity scores for each parameter size constraint. QAT results (bottom) are averaged over 25 runs, each conducted for 5 epochs with 25,000 samples. LIMPQ source code does not provide search configurations for ResNet34.

for fine-tuning. As shown in Figure 4, we observe that bit-width assignments from *InfoQ* also lead to higher accuracy after a short QAT schedule. This confirms that the benefits of our more accurate sensitivity metric compound during the fine-tuning process.

Search Efficiency Analysis

A key advantage of *InfoQ* is its high search efficiency. The cost is composed of two phases: a one-time sensitivity analysis and the near-instantaneous bit allocation.

Sensitivity Analysis Cost. The primary computational cost of our method is a one-time, training-free analysis. For a network with L quantizable layers and a set of candidate bit-widths $\boldsymbol{B}$, the process requires $L \times |\boldsymbol{B}|$ forward passes over a small, labeled calibration dataset to collect the necessary activations. From these activations, we compute SMI estimates at the pre-selected observer layers. This entire analysis is backpropagation-free and highly parallelizable.

The selection of observer layers is performed via a correlation analysis on a small subset of the calibration data. Importantly, the values from this step can be reused in the subsequent sensitivity analysis (Algorithm 1, Eq. 4), making its computational cost negligible.

Retraining-based proxy methods like LIMPQ (Tang et al. 2023) require extensive training on a large fraction of the dataset (e.g., 600,000 samples for ImageNet) to learn their sensitivity indicators. In contrast, Hessian-based methods like HAWQv2 are faster but, as we have shown, yield sub-optimal bit assignments due to their local scope. *InfoQ* thus occupies a unique and highly favorable position in the accuracy-efficiency trade-off, achieving SOTA accuracy with a practical, one-shot analysis, as shown in the Figure 3.

Bit Allocation Cost. Once the sensitivity scores $S(\ell, b)$ are pre-computed, the bit allocation phase is extremely fast. For any given hardware constraint (e.g., a target model size or BitOps), the optimal bit-width configuration is found by solving the ILP problem (Eq. 5), by an off-the-shelf solver in milliseconds. A significant practical benefit of this decoupling is that the same set of pre-computed sensitivity scores can be reused instantly to generate optimal configurations for multiple different hardware targets or budget constraints, without re-running the more expensive sensitivity analysis.

6 Conclusion

In this work, we addressed a fundamental limitation of existing criterion-based MPQ methods: their reliance on local sensitivity metrics. We argued that the true impact of quantizing a layer is a global phenomenon, best measured by the resulting disruption to the network information flow. To this end, we introduced *InfoQ*, a novel framework that is based on this principle. By using Sliced Mutual Information to directly quantify the cascading effects of quantization on downstream layers, *InfoQ* constructs an accurate, global sensitivity metric. This metric enables the determination of optimal bit-width configurations through a fast, training-free analysis, followed by an instantaneous ILP-based allocation.

Our extensive experiments on ImageNet demonstrate that this information-theoretic approach yields SotA results, outperforming both Hessian-based and retraining-based methods on ResNet and MobileNetV2 architectures, particularly in high-compression regimes. Our work establishes a practical methodology for applying information-theoretic diagnostics to model compression, providing a principled foundation for bit-width (and more in general, perturbation robustness) allocation in deep neural network layers.

Acknowledgments

This paper is supported by the PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

References

Amjad, R. A.; and Geiger, B. C. 2018. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42: 2225–2239.

Baskin, C.; Zheltonozhkii, E.; Rozen, T.; Liss, N.; Chai, Y.; Schwartz, E.; Giryes, R.; Bronstein, A. M.; and Mendelson, A. 2021. NICE: Noise Injection and Clamping Estimation for Neural Network Quantization. *Mathematics*, 9(17): 2144.

Butakov, I.; Tolmachev, A. D.; Malanchuk, S.; Neopryatnaya, A. M.; Frolov, A. A.; and Andreev, K. V. 2023. Information Bottleneck Analysis of Deep Neural Networks via Lossy Compression. *ArXiv*, abs/2305.08013.

Cai, Z.; and Vasconcelos, N. 2020. Rethinking Differentiable Search for Mixed-Precision Neural Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2346–2355.

Chen, W.; Wang, P.; and Cheng, J. 2021. Towards Mixed-Precision Quantization of Neural Networks via Constrained Optimization. arXiv:2110.06554.

Choi, J.; Wang, Z.; Venkataramani, S.; Chuang, P. I.-J.; Srinivasan, V.; and Gopalakrishnan, K. 2018. PACT: Parameterized Clipping Activation for Quantized Neural Networks. arXiv:1805.06085.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, Z.; Wang, X.; Sharify, S.; and Orshansky, M. 2023. Mixed-Precision Quantization for Deep Vision Models with Integer Quadratic Programming.

Dong, P.; Li, L.; Wei, Z.; Niu, X.-Y.; Tian, Z.; and Pan, H. 2023. EMQ: Evolving Training-free Proxies for Automated Mixed Precision Quantization. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17030–17040.

Dong, Z.; Yao, Z.; Cai, Y.; Arfeen, D.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2019a. HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks. arXiv:1911.03852.

Dong, Z.; Yao, Z.; Gholami, A.; Mahoney, M.; and Keutzer, K. 2019b. HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision. arXiv:1905.03696.

Défossez, A.; Adi, Y.; and Synnaeve, G. 2022. Differentiable Model Compression via Pseudo Quantization Noise. arXiv:2104.09987.

Esser, S. K.; McKinstry, J. L.; Bablani, D.; Appuswamy, R.; and Modha, D. S. 2019. Learned Step Size Quantization. *ArXiv*, abs/1902.08153.

Geiger, B. C. 2020. On Information Plane Analyses of Neural Network Classifiers—A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 33: 7039–7051.

Goldfeld, Z.; Greenewald, K. H.; Niles-Weed, J.; and Polyanskiy, Y. 2019. Convergence of Smoothed Empirical Measures With Applications to Entropy Estimation. *IEEE Transactions on Information Theory*, 66: 4368–4391.

Goldfeld, Z.; Greenewald, K. H.; Nuradha, T.; and Reeves, G. 2022. k-Sliced Mutual Information: A Quantitative Study of Scalability with Dimension. In *Neural Information Processing Systems*.

Guo, S.; Zhang, L.; Zheng, X.; Wang, Y.; Li, Y.; Chao, F.; Wu, C.; Zhang, S.; and Ji, R. 2023. Automatic Network Pruning via Hilbert-Schmidt Independence Criterion Lasso under Information Bottleneck Principle. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17412–17423.

Guo, Z.; Zhang, X.; Mu, H.; Heng, W.; Liu, Z.; Wei, Y.; and Sun, J. 2020. Single Path One-Shot Neural Architecture Search with Uniform Sampling. arXiv:1904.00420.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. *arXiv: Computer Vision and Pattern Recognition*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2003. Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69 6 Pt 2: 066138.

Lorenzen, S. S.; Igel, C.; and Nielsen, M. 2021. Information Bottleneck: Exact Analysis of (Quantized) Neural Networks. *ArXiv*, abs/2106.12912.

Lou, Q.; Guo, F.; Liu, L.; Kim, M.; and Jiang, L. 2020. AutoQ: Automated Kernel-Wise Neural Network Quantization. arXiv:1902.05690.

Ma, Y.; Jin, T.; Zheng, X.; Wang, Y.; Li, H.; Jiang, G.; Zhang, W.; and Ji, R. 2021. OMPQ: Orthogonal Mixed Precision Quantization. *ArXiv*, abs/2109.07865.

Mitchell, S.; O'Sullivan, M.; and Dunning, I. 2011. PuLP : A Linear Programming Toolkit for Python.

Nielsen, M. Ø.; Østergaard, J.; Jensen, J. H.; and Tan, Z.-H. 2021. Compression of DNNs Using Magnitude Pruning and Nonlinear Information Bottleneck Training. *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. Q.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y. B.; Li, S.-W.; Misra, I.; Rabbat, M. G.; Sharma, V.; Synnaeve, G.; Xu, H.; Jégou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. *ArXiv*, abs/2304.07193.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381.

Tang, C.; Ouyang, K.; Wang, Z.; Zhu, Y.; Wang, Y.; Ji, W.; and Zhu, W. 2023. Mixed-Precision Neural Network Quantization via Learned Layer-wise Importance. arXiv:2203.08368.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, 1–5.

Uhlich, S.; Mauch, L.; Cardinaux, F.; Yoshiyama, K.; García, J. A.; Tiedemann, S.; Kemp, T.; and Nakamura, A. 2019. Mixed Precision DNNs: All you need is a good parametrization. In *International Conference on Learning Representations*.

Wang, K.; Liu, Z.; Lin, Y.; Lin, J.; and Han, S. 2019. HAQ: Hardware-Aware Automated Quantization with Mixed Precision. arXiv:1811.08886.

Wu, B.; Wang, Y.; Zhang, P.; Tian, Y.; Vajda, P.; and Keutzer, K. 2018. Mixed Precision Quantization of ConvNets via Differentiable Neural Architecture Search. arXiv:1812.00090.

Yang, L.; and Jin, Q. 2020. FracBits: Mixed Precision Quantization via Fractional Bit-Widths. In *AAAI Conference on Artificial Intelligence*.

Yu, H.; Han, Q.; Li, J.; Shi, J.; Cheng, G.; and Fan, B. 2020. Search What You Want: Barrier Panelty NAS for Mixed Precision Quantization. arXiv:2007.10026.

Zhang, D.; Yang, J.; Ye, D.; and Hua, G. 2018. LQ-Nets: Learned Quantization for Highly Accurate and Compact Deep Neural Networks. arXiv:1807.10029.

Zheng, X.; Ma, Y.; Xi, T.; Zhang, G.; Ding, E.; Li, Y.; Chen, J.; Tian, Y.; and Ji, R. 2021. An Information Theory-inspired Strategy for Automatic Network Pruning. *ArXiv*, abs/2108.08532.