ROTATION INVARIANT QUANTIZATION FOR MODEL COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training Neural Network (NN) model compression is an attractive approach for deploying large, memory-consuming models on devices with limited memory resources. In this study, we investigate the theoretical limits of NN model compression using rate-distortion theory. First, we suggest a Rotation-Invariant Quantization (RIQ) technique that utilizes a single parameter to quantize the entire NN model, yielding a different rate at each layer, i.e., mixed-precision quantization. Then, we prove that our rotation-invariant approach is optimal in terms of compression. We rigorously evaluate RIQ and demonstrate its capabilities on various models and tasks. For example, RIQ facilitates $\times 19.4$ and $\times 52.9$ compression ratios on pre-trained VGG dense and pruned models, respectively, with $< 0.4\%$ accuracy degradation. The code is available in the supplementary material.

1 INTRODUCTION

Deep neural networks are widely used for various tasks, such as computer vision, *Natural Language Processing* (NLP), and recommendation systems. Nevertheless, while performance continuously improves, the models become larger with a massive increase in the number of parameters. In fact, modern *Neural Network* (NN) models may have billions and even trillions of parameters, which makes the deployment of these models a challenging task (Chang et al., 2020). One way to mitigate this issue is compressing the model's parameters to reduce its overall memory footprint while satisfying an accuracy constraint. Namely, obtaining a smaller model that is (almost) as capable as the original model.

The most common model compression techniques are weight pruning, quantization, knowledge distillation, and low-rank decomposition. Such optimizations strive to find a smaller model while keeping the original model's accuracy, overlooking the potential inherent in its entropy limit. In the context of NN models, the entropy value is of particular interest as it provides the (theoretical) number of bits required for representing the model parameters. The optimal compression asymptotically attains this entropy limit.

In this context, lossy compression gives considerable merit as it facilitates reducing the NN size significantly with negligible accuracy degradation. The key steps of this approach are the *quantization* and the *compression* (Polyanskiy & Wu, 2014, Ch. 25). In the quantization phase, the number of unique weight values (symbols) is reduced, consequently reducing the model's entropy. Modifying the symbols' statistics, however, introduces distortion (i.e., quantization error) with respect to the original model. Hence, proper quantization methods are substantial as they determine both the resulting entropy and the distortion of the model's output.

In the compression phase, redundant information is removed, reaching the most compact representation possible without introducing further errors. Thus, the only distortion originates from altering the symbols' statistics, during the quantization. Finding a solution that quantizes the model at the lowest possible bit rate while satisfying a certain distortion requirement is at the heart of quantization optimization problems, and is known as the rate-distortion problem (Cover & Thomas, 2006).

In this study, we investigate the theoretical limits of NN model compression using rate-distortion theory, where the distortion is measured by a cosine distance between the outputs of the original and the quantized model. In particular, we formulate the model compression as an optimization problem, where the goal is maximizing the model compression ratio subject to a given distortion constraint.

Our focus is *mixed-precision solutions*, where each layer gets quantized at a possibly different rate. Specifically, the main contribution is as follows.

- We design a post-training Rotation-Invariant Quantization (RIQ) method that quantizes the entire model subject to a distortion constraint as a function of a single parameter. The main theme of our approach is picking the quantization bin width to be proportional to the layers' norm. Consequently, considering a vector representation of the weights in each layer, the resulting distortion is indifferent to the orientation of the layers. To find the optimal quantization parameter efficiently, we suggest a searching paradigm that bounds the search space and performs nested refinements on this space. This approach minimizes the per-layer entropy and thus optimizes the resulting compression.
- To analyze the optimality of the RIQ algorithm, we introduce a surrogate model that depicts quantization in terms of rotation of the model weights. Its analysis reveals that the rate-distortion minimizing distribution for NN models is a spherical (rotation invariant) distribution constructed by the product of layers' spherical distribution. Due to convexity, the rate achieved under this product distribution is bounded by a rate achieved under the layers' average spherical distribution. Further, to optimize the latter, a single quantization parameter suffices. Consequently, the RIQ follows these guidelines, i.e., being indifferent to rotations and using a single parameter to optimize the quantization of the entire model.
- We rigorously evaluate the RIQ and demonstrate its capabilities on various models and tasks. RIQ attains a remarkable compression ratio with a negligible accuracy loss in all assessments, surpassing recent results in the literature.

2 Related Work

This section is devoted to prior work on model compression that is most relevant to this study. Roughly speaking, typical model compression methods can be classified into four categories. Weight pruning, quantization, knowledge distillation (Sarfraz et al., 2021; Walawalkar et al., 2020), and low-rank decomposition (Idelbayev & Carreira-Perpinán, 2020; Lin et al., 2018; Lee et al., 2019). Even though such methods strive to find a smaller model while retaining the model's accuracy, they often tend to neglect the potential inherent in the entropy limit. In this study, we seek to minimize the model entropy by quantization and then attain this entropy limit by compression while satisfying a distortion requirement at the model's output.

Quantization is a prominent method for compressing NN models. In Wu et al. (2020); Banner et al. (2019), the authors considered fixed-bit quantization methods, where all layers are quantized at the same integer bit rate. In this paper, on the other hand, we consider post-training mixed-precision solutions. Bhalgat et al. (2020); Wang et al. (2019); Idelbayev et al. (2021) focused on quantization-aware training, where the weights quantization is performed during the training. To attain lower quantization rates, Fan et al. (2020); Baskin et al. (2021) suggested training the models with noise. Although quantization-aware training methods may achieve better results than the post-training approaches, they are time-consuming, and thus, may not be applicable for deployment purposes. Cai et al. (2020); Hubara et al. (2021) proposed knowledge distillation techniques to minimize Kullback–Leibler (KL) divergence between the layers of the quantized and original model, using a small calibration set. Nevertheless, these works determine the quantization bin width by optimizing the quantization range, which depends on the orientation of the weights. Consequently, the bin width is sensitive to rotations. On the other hand, RIQ designs the bin width in proportion to the norm of the weights, and thus, the bin width is indifferent to rotations.

The idea of pruning NN connections based on information-theoretic ideas was explored already in the seminal works (LeCun et al., 1989; Hassibi et al., 1993). Later, Han et al. (2015) used magnitude threshold to remove redundant parameters, and then, utilized Huffman's entropy coding to compress these parameters. Since then, pruning techniques gained popularity, searching for effective methods to prune parameters (Zhang et al., 2021; Frankle & Carbin, 2018; Lee et al., 2019). For assessing pruning and quantization algorithms, Gao et al. (2019); Isik et al. (2022) provided a rate-distortion theory framework, showing that entropy reduction during training is beneficial, as low-entropy models are more amenable to compression (Oktay et al., 2019; Baskin et al., 2019). This study continues these guidelines, using rate-distortion theory to analyze rotation-invariant solutions and further provides enhancements that find the optimal solution efficiently.

3 PRELIMINARIES

In this section, we rigorously define the model compression optimization problem and the relevant known results on quantization and the rate-distortion theory. Throughout, w and $\mathbf{w}$ (bold) denote scalar variable and vector, respectively, unless stated otherwise. $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denotes the standard ℓ^2-norm and the inner product, respectively. We use $p_{\mathrm{w}}(w)$ to denote the probability distribution of a random variable w. Hereafter, $\mathbf{w}_{[1:L]} = \{\mathbf{w}_1, ..., \mathbf{w}_L\} \in \mathbb{R}^{n_1 \times \cdots \times n_L}$ denotes the weights of a pretrained model with L layers, where $\mathbf{w}_\ell \in \mathbb{R}^{n_\ell}$ are n_ℓ weights of layer ℓ. The quantized representation of those weights is denoted by $\hat{\mathbf{w}}_\ell$.

3.1 PROBLEM STATEMENT

Given a *pretrained model* f that characterizes a prediction of input space $\mathcal{X}$ by L layers, where each layer ℓ is parameterized with n_ℓ weights $\mathbf{w}_\ell \in \mathbb{R}^{n_\ell}$, our goal is to obtain the smallest (quantized and compressed) version of this model $\hat{f}$, whose output is as close as possible to the output of f. To assess the fidelity of the quantization, a sample $\mathbf{x}$ is sent through f and $\hat{f}$, and the *distortion between the outputs* $d(f(\mathbf{x}), \hat{f}(\mathbf{x}))$ is measured. In this study, we focus on the cosine distance as distortion measure of the outputs. That is,

$$d(f(\mathbf{x}), \hat{f}(\mathbf{x})) \triangleq 1 - \frac{\langle f(\mathbf{x}), \hat{f}(\mathbf{x})\rangle}{\|f(\mathbf{x})\| \cdot \|\hat{f}(\mathbf{x})\|} \tag{1}$$

This distortion measures the rotation angle that is required to align $\hat{f}(\mathbf{x})$ with $f(\mathbf{x})$, and according to our experiments, it serves as a decent proxy to various scoring methods such as accuracy and perplexity, (Usino et al., 2019; Maharani et al., 2020).

One may describe the linear part (e.g., matrix multiplication) in each layer by rotation and scale, where the weights rotate the input. To provide intuition and motivation for rotation-invariant quantization of the weights, however, we describe the opposite direction (i.e., the input rotates the weights in each layer). As $\mathbf{x}$ passes the first layer of f and $\hat{f}$, it rotates (and scales) $\mathbf{w}_1$ and $\hat{\mathbf{w}}_1$, respectively. Due to the quantization error, $\mathbf{x}$ acts differently on $\mathbf{w}_1$ and $\hat{\mathbf{w}}_1$, and thus, yields unequal outputs. These unequal outputs rotate the next layer's weights, and so on, reaching the output of f and $\hat{f}$. The resulting distortion in eq. (1) relates to the quantization errors gathered through the layers, essentially reflected as rotation divergence. Accordingly, each quantized layer produces a rotation error in its output, and this error keeps propagating and accumulating through the layers until reaching the model's output. Clearly, different samples induce different distortions. Yet, if the quantized weights are indifferent to rotations, the resulting distortion would be similar for different samples. Since we wish to minimize the quantized model's empirical entropy while satisfying a certain distortion requirement, this becomes an attractive approach.

Formally, given a trained model f, and a sample $\mathbf{x}$, we wish to find a quantized model $\hat{f}$ whose weights $\hat{\mathbf{w}}_{[1:L]}$ solves the following optimization problem.

$$\begin{aligned} \underset{\hat{f}}{\text{minimize}} \quad & H(\hat{\mathbf{w}}_{[1:L]}) \\ \text{subject to} \quad & d(f(\mathbf{x}), \hat{f}(\mathbf{x})) \leq D \end{aligned}$$

for some distortion requirement D, where $H(\hat{\mathbf{w}}_{[1:L]})$ is the empirical entropy function defined under the empirical probability of a collection of random variables (Cover & Thomas, 2006, eq. (2.48) and Ch. 7.6). In this work, we characterize the properties of the minimizing $\hat{f}$ and devise a searching method that finds the minimizing $\hat{f}$ efficiently.

In this formulation, we consider *mixed-precision quantization* solutions, where the weights $\mathbf{w}_\ell$ of each layer ℓ are quantized at a different rate R_ℓ. That is, the average number of bits per symbol, R_ℓ, varies over the layers. Moreover, the rate values R_ℓ are not limited to integer values as typically considered. Allowing *non-integer quantization rates* may seem peculiar at first sight, since the number of possible symbols for an integer number of bits is a power of 2, and thus, using fewer symbols than this means that those bits are not utilized well. However, using fewer symbols is often beneficial in terms of the entropy of the quantized weights. Specifically, to obtain the smallest model,

we must pick a solution whose entropy is minimal. This entropy can be achieved by compressing the layers' weights of the model $\hat{f}$ with an entropy achieving encoder, e.g., an arithmetic encoder such as the *Asymmetric Numeral Systems* (ANS) (Duda, 2013). The resulting compression ratio, assuming 32 bits representation of the source symbols, is approximately

$$\text{Compression Ratio} = \frac{32 \cdot \sum_{\ell=1}^{L} n_\ell}{\sum_{\ell=1}^{L} n_\ell \cdot H(\hat{\mathbf{w}}_\ell) + |T_\ell|} \tag{2}$$

Where $|T_\ell|$ denotes the coding table size of layer ℓ.

3.2 Rate-Distortion Theory

The rate-distortion theory determines the *minimum number of bits per symbol*, or simply the minimum bit rate, required for describing a random variable with a certain (average) distortion. In particular, to quantize a sequence of n independent realizations $\boldsymbol{w} = (w_1, \cdots, w_n)$, generated by a source $\mathbf{w}$ with distribution $p_{\mathbf{w}}(\boldsymbol{w}), \boldsymbol{w} \in \mathcal{W}^n$ into R bits, encoding and decoding functions are utilized. The encoder $\mathcal{E} : \mathcal{W}^n \to \{0,1\}^R$ maps the sequence to one of 2^R possible indices, and the decoder $\mathcal{D} : \{0,1\}^R \to \hat{\mathcal{W}}^n$ maps the given index into an estimated (quantized) sequence $\hat{\boldsymbol{w}} = (\hat{w}_1, \cdots, \hat{w}_n)$. Thus, the rate-distortion pair (R, D) are the resulting rate R and distance $D = d(\boldsymbol{w}, \hat{\boldsymbol{w}})$ between the original sequence and the quantized sequence.

In general, we wish to minimize both the rate and the distortion, however, there is an inherent tradeoff, characterized by the rate-distortion function as (Cover & Thomas, 2006, Ch. 10)

$$R(D) = \min_{p(\hat{\boldsymbol{w}}|\boldsymbol{w}):\mathbb{E}[d(\boldsymbol{w},\hat{\boldsymbol{w}}))]\leq D} I(\mathbf{w}; \hat{\mathbf{w}}) \tag{3}$$

where $I(\mathbf{w}; \hat{\mathbf{w}}) = H(\mathbf{w}) - H(\mathbf{w}|\hat{\mathbf{w}})$ is the mutual information between the source vector $\mathbf{w}$ and its reconstruction $\hat{\mathbf{w}}$ (Cover & Thomas, 2006, Ch. 2.4), and $d(\cdot, \cdot)$ is a predefined distortion metric, such as the cosine distance in eq. (1). Thus, the rate-distortion function determines the infimum rate R that achieves a given distortion D. This infimum is attained by minimizing overall conditional distributions $p(\hat{\boldsymbol{w}}|\boldsymbol{w})$ for which distortion D is satisfied under the considered $p(\boldsymbol{w})$.

3.3 Uniform Scalar Quantization

The rate-distortion theory tells that it is optimal to describe the whole sequence jointly, using one of 2^R indices, even when the variables are i.i.d. Yet, in terms of entropy, Koshelev (1963) showed that uniform scalar quantization is (asymptotically) optimal when one intends to further compress the quantized data losslessly. Since this paper considers the latter approach, this section briefly discusses uniform scalar quantization and its analysis.

For a random variable $\mathrm{w} \in [-A/2, A/2]$, where $A \in \mathbb{R}$, uniform quantization partitions the range $[-A/2, A/2]$ into N bins uniformly, such that each bin has width $\Delta = A/N$. Thus, any realization of w is encoded (rounded) into *an integer value*, $\lceil \mathrm{w}/\Delta \rfloor$, that corresponds to its bin index. The decoder then reconstructs its value by

$$\hat{\mathrm{w}} = \lceil \mathrm{w}/\Delta \rfloor \cdot \Delta \tag{4}$$

The fidelity of this quantization is typically measured by a distortion measure, such as the *Mean Squared Error* (MSE) criterion, defined as $D(N) = \mathbb{E}|\mathrm{w} - \hat{\mathrm{w}}|^2$. To analyze, it is more convenient to examine the quantization in terms of rate $R = \log_2 N$. In high-rate regime (i.e., $R \gg 1$), the *probability density* in each bin is nearly flat, and consequently, the expected distortion is (Polyanskiy & Wu, 2014, Ch. 25.1)

$$\mathbb{E}|\mathrm{w} - \hat{\mathrm{w}}|^2 = \Delta^2/12 \tag{5}$$

Further, the resulting entropy of the quantized symbol is (Cover & Thomas, 2006, Theorem 8.3.1)

$$H(\hat{\mathrm{w}}) = H(\mathrm{w}) - \log(\Delta) \quad \text{[bits/symbol]} \tag{6}$$

where $H(\mathrm{w}) = -\sum_{w \in \mathcal{W}} p(w) \log p(w)$ is the entropy function (Cover & Thomas, 2006, eq. (2.1)). In other words, quantization reduces the entropy by $\log \Delta$, and thus, a larger Δ yields a lower entropy, and hence, potentially, a higher compression ratio by eq. (2).

4 ROTATION-INVARIANT MIXED-PRECISION QUANTIZATION

In this section, we present the RIQ method, which yields a different quantization rate in each layer (i.e., mixed-precision solution) while satisfying the distortion requirement in eq. (1). Then, we use the rate-distortion theory to analyze its performance. The analysis reveals that the problem simplifies to a single (average) layer optimization, which is driven by a single quantization parameter.

4.1 THE RIQ ALGORITHM

The key motivation for rotation-invariant quantization is being indifferent to the orientation of the weights $\mathbf{w}_\ell$. Specifically, picking a bin width Δ_ℓ that remains constant under any rotation of $\mathbf{w}_\ell$. Typical quantizations, on the other hand, determine Δ_ℓ according to the range $\max(\mathbf{w}_\ell)-\min(\mathbf{w}_\ell)$, whose value depends on the orientation of $\mathbf{w}_\ell$, and hence, Δ_ℓ changes as $\mathbf{w}_\ell$ is being rotated.

Accordingly, RIQ designs the bin width, Δ_ℓ, in proportion to the length $\|\mathbf{w}_\ell\|$ in each layer. Since length is invariant to rotations, the resulting Δ_ℓ is indifferent to the orientation of $\mathbf{w}_\ell$. Consequently, the resulting distortion is also indifferent to the orientation of $\mathbf{w}_\ell$, as the bin width dictates the distortion. To examine the relation between Δ_ℓ, the length $\|\mathbf{w}_\ell\|$, and the resulting distortion $d(\mathbf{w}_\ell, \hat{\mathbf{w}}_\ell)$ for each layer ℓ, the following lemma is useful.

Lemma 1. *Let $\epsilon_\ell \triangleq d(\mathbf{w}_\ell, \hat{\mathbf{w}}_\ell)$ be the distortion of layer ℓ, then, the quantization bin width satisfies*

$$\Delta_\ell = \sqrt{\epsilon_\ell} \cdot \|\mathbf{w}_\ell\| \cdot \sqrt{24/n_\ell}$$

The proof of Lemma 1 is elaborated in Appendix A.1. Note that when $\sqrt{\epsilon_\ell} = \frac{1}{k}\sqrt{\frac{n_\ell}{24}}$, where k is a parameter to be optimized, the resulting bin width is $\Delta_\ell(k) = \|\mathbf{w}_\ell\|/k$, which is proportional to the layer's length as desired. Further, note that $\Delta_\ell(k)$ and ϵ_ℓ are monotonically decreasing as k increases. Consequently, by eq. (6), the entropy increases with k. Thus, we seek the smallest k (i.e., minimum entropy) solution that satisfies the distortion requirement.

Note that as k increases, $\Delta_\ell(k) \to 0$. To prevent this, we add a small constant ϵ_0 to $\sqrt{\epsilon_\ell}$, which bounds the value of the smallest $\Delta_\ell(k)$. In this case, setting $\sqrt{\epsilon_\ell} = \frac{1}{k}\sqrt{\frac{n_\ell}{24}} + \epsilon_0$, yields,

$$\Delta_\ell(k) = \|\mathbf{w}_\ell\| \cdot \left(\frac{1}{k} + \epsilon_0 \cdot \sqrt{\frac{24}{n_\ell}}\right) \tag{7}$$

Still, being rotation-invariant per layer does not guarantee to be rotation-invariant model-wise. However, when using a single parameter k for the whole model, it automatically preserves the proportion between the set of $\Delta_{[1:L]}$. In Section 4.2, we prove that the optimal solution is indeed a model-wise rotation invariant solution.

The objective of RIQ is is efficiently finding the optimal k parameter which satisfies the distortion requirement in eq. (1). Accordingly, we introduce an efficient iterative searching algorithm for finding the optimal k. In each iteration, the algorithm refines the searching range until reaching the smallest k (up to a small constant) that satisfies the distortion requirement.

Since optimizing k over unbounded search space is intricate and time-consuming, we derive searching bounds for the optimal k in the following proposition. Let k^* be the optimal (smallest) k that satisfies constraint D, and let ℓ^* be the index of the layer with the largest n_ℓ in f.

Proposition 1. *The optimal k^* satisfies the following bounds: $\frac{\sqrt{n_{\ell^*}/24}}{(1-\epsilon_0)} \leq k^* \leq \frac{\sqrt{n_{\ell^*}/24}}{(\epsilon_0 \cdot \sqrt{\epsilon_0})}$.*

The proof is deferred to Appendix A.2. To further improve the search time, a nested refinement is utilized. Specifically, at each stage, only $\sqrt{|O(k)|}$ values of k in ascending order are considered. Once a certain value of k satisfies the distortion requirement D, this k becomes the new upper limit for the search, and the search region is refined within a smaller region of k to consider, again with only $\sqrt{|O(k)|}$ potential values to inspect. This repeatedly continues until the search step is sufficiently small. These refinements enable fast convergence in relatively few iterations. See Algorithm 1 for a detailed description of RIQ.

Algorithm 1: The RIQ algorithm

Data: model weights $\mathbf{w}_{[1:L]}$, distortion requirement D, minimum error ϵ_0

Result: quantized weights $\hat{\mathbf{w}}_{[1:L]}$, such that $d(f(\mathbf{x}), \hat{f}(\mathbf{x})) \leq D$

Initialize: $k_{\min} = \frac{\sqrt{n_{\ell^*}/24}}{1-\epsilon_0}$, $k_{\max} = \frac{\sqrt{n_{\ell^*}/24}}{\sqrt{\epsilon_0}\cdot\epsilon_0}$, $k = k_{\min}$, step $= \sqrt{k_{\max} - k_{\min}}$;

while $k \leq k_{\max}$ **do**

  for $\ell = 1, \cdots, L$ **do**

    $\Delta_\ell = \|\mathbf{w}_\ell\| \cdot \left(\frac{1}{k} + \epsilon_0 \cdot \sqrt{\frac{24}{n_\ell}}\right)$;

    $\hat{\mathbf{w}}_\ell = \left\lceil \frac{\mathbf{w}_\ell}{\Delta_\ell} \right\rfloor \cdot \Delta_\ell$;

  end

  if $d(f(\mathbf{x}), \hat{f}(\mathbf{x})) \leq D$ **then**

    if step ≤ 3 ; `/* Stop condition */`

    then

      compress to $H(\hat{\mathbf{w}}_{[1:L]})$ with entropy achieving encoder;

    else

      $k_{\max} = k$;

      step $= \sqrt{\text{step}}$;

      $k = k - \text{step} \cdot \lfloor \text{step} \rfloor$

    end

  else

    $k = k + \text{step}$

  end

end

Remark. *The additional degree of freedom that ϵ_0 gives is substantial. For example, it facilitates enforcing quantization to R bits (e.g., $R = 8$ bits) for low precision runtime, by setting the limit $k \to \infty$ at $\epsilon_0(\ell) = \frac{\max(\mathbf{w}_\ell) - \min(\mathbf{w}_\ell)}{2^R - 1} / \sqrt{\frac{24 \cdot \|\mathbf{w}_\ell\|^2}{n_\ell}}$.*

If one wishes to minimize the difference between the histogram of $\hat{\mathbf{w}}_\ell$ and the theoretical probability distribution of $\mathbf{w}_\ell$ at the limit $k \to \infty$, then, by the Freedman & Diaconis (1981) rule, one may set $\epsilon_0(\ell) = 2\frac{IQR(\mathbf{w}_\ell)}{\sqrt[3]{n_\ell}}$, where $IQR(\mathbf{w}_\ell)$ is the inter-quartile range of $\mathbf{w}_\ell$.

For simplicity, in the sequel we apply the same small common constant value ϵ_0 to all layers.

Remark. *The proof of Lemma 1 may serve as a proxy to other interesting error criteria such as the MSE and the* Signal to Quantization Noise Ratio *(SQNR) (Caffarena et al., 2010). Clearly, this yields another relation between Δ_ℓ and ϵ_ℓ. Details in the Appendix A.5.*

4.2 RIQ Rate-Distortion Analysis

In this section, we provide theoretical justification for the optimality of RIQ. We introduce a surrogate model for which the rate-distortion analysis for NN models is tractable, showing that the quantization error (distortion) under the minimizing distribution is indifferent to the orientation of $\mathbf{w}_\ell$, as RIQ suggests.

First, extending eq. (4) to NN model quantization, where layer ℓ is encoded uniformly, yields

$$\hat{\mathbf{w}}_\ell = \lceil \mathbf{w}_\ell / \Delta_\ell \rfloor \cdot \Delta_\ell \tag{8}$$

To analyze the rate-distortion of eq. (8) for the cosine distance in eq. (1), it is essential to consider quantization as *rotation and scale* in each layer. Accordingly, we define the following surrogate model.

Surrogate Model. *Let $\mathbf{w}_\ell$ be the weights of layer ℓ, and let $\hat{\mathbf{w}}_\ell$ denote their quantized representation. Let θ_ℓ be the rotation angle from $\mathbf{w}_\ell$ to $\hat{\mathbf{w}}_\ell$, such that $\langle \frac{\mathbf{w}_\ell}{\|\mathbf{w}_\ell\|}, \frac{\hat{\mathbf{w}}_\ell}{\|\hat{\mathbf{w}}_\ell\|} \rangle = \cos(\theta_\ell)$, and let $\mathbf{U}(\theta_\ell | \mathbf{w}_\ell) \in \mathbb{R}^{n_\ell \times n_\ell}$ be a random orthogonal transformation corresponding to a random rotation that is θ_ℓ away from $\mathbf{w}_\ell$. Then,*

$$\tilde{\mathbf{w}}_\ell = \|\hat{\mathbf{w}}_\ell\| \cdot \mathbf{U}(\theta_\ell | \mathbf{w}_\ell) \frac{\mathbf{w}_\ell}{\|\mathbf{w}_\ell\|} \tag{9}$$

models the quantized weights $\hat{\mathbf{w}}_\ell$.

Intuitively, $\mathbf{U}(\theta_\ell|\mathbf{w}_\ell)$ randomly rotates any given vector uniformly on a sphere, where one degree of freedom is lost due to the requirement of being θ_ℓ away from $\mathbf{w}_\ell$. To obtain $\hat{\mathbf{w}}_\ell$ in eq. (8), the realization of $\mathbf{U}(\theta_\ell|\mathbf{w}_\ell)$ should rotate the unit vector $\mathbf{w}_\ell/\|\mathbf{w}_\ell\|$ in the plane generated by $\mathbf{w}_\ell$ and $\hat{\mathbf{w}}_\ell$, and then, stretches it into the length $\|\hat{\mathbf{w}}_\ell\|$. In other words, this model describes a random vector $\tilde{\mathbf{w}}_\ell$ that is uniformly distributed on a cone that is θ_ℓ away from $\mathbf{w}_\ell$, for which a single realization matches eq. (8). The merit of this model is its tractable analysis, from which *spherically symmetric distribution* emerges to depict the quantized weights (Fang et al., 2018, Definition 2.1).

Definition. *A random vector* $\mathbf{z} \in \mathbb{R}^n$ *is said to have a spherical (rotation-invariant) distribution if its distribution does not change under any orthogonal transformation* $\mathbf{U} \in \mathbb{R}^{n\times n}$*, i.e.,*

$$p(\mathbf{z}) \overset{d}{=} p(\mathbf{U}\mathbf{z}).$$

The following proposition characterizes the indifference of $\tilde{\mathbf{w}}_\ell$ to the orientation of a given $\mathbf{w}_\ell$.

Proposition 2. *Let* $\mathbf{w}_\ell$ *be the weights of layer* ℓ*, and let* $\tilde{\mathbf{w}}_\ell$ *model the quantized representation of those weights, modeled by eq.* (9)*. Then,* $\tilde{\mathbf{w}}_\ell|\mathbf{w}_\ell$ *have a spherical (rotation-invariant) distribution.*

A detailed proof is given in Appendix A.3. Essentially, the strength of Proposition 2 is twofold. First, it proves that the distribution of $\tilde{\mathbf{w}}_\ell|\mathbf{w}_\ell$ in each layer does not change when arbitrary rotations are applied to it. Second, it holds for any distribution of $\mathbf{w}_\ell$ and θ_ℓ. This further indicates that the quantization error reflected as the deviation θ_ℓ is not affected by such rotations. The following theorem extends the results of Proposition 2 to multiple layers, showing that spherical distribution is also the minimizing distribution of the rate-distortion problem of NN models.

Theorem 1. *Let* $f(\cdot)$ *be a NN model with* L *layers whose weights are* $\mathbf{w}_{[1:L]}$*, and let* $\tilde{\mathbf{w}}_{[1:L]}$ *model the quantized representation of those weights. Then, the unique minimizing distribution* $p(\tilde{\boldsymbol{w}}_{[1:L]}|\boldsymbol{w}_{[1:L]})$ *of the rate-distortion function*

$$R(D) = \min_{\substack{p\left(\tilde{\boldsymbol{w}}_{[1:L]}\middle|\boldsymbol{w}_{[1:L]}\right):\\ d(f(\mathbf{x}),\hat{f}(\mathbf{x}))\le D}} I\left(\mathbf{w}_{[1:L]};\tilde{\mathbf{w}}_{[1:L]}\right) \tag{10}$$

is a product distribution constructed as the product of the layers' spherical distribution.

The detailed proof is given in Appendix A.4. The key steps of the proof are, first, showing that the minimizing distribution is a product distribution. Then, due to the convexity of the rate-distortion in eq. (10), we bound the mutual information with the average distribution of θ_ℓ taken over the model's layers. Consequently, the problem is simplified to a single (average) layer optimization, which is governed by a single quantization parameter.

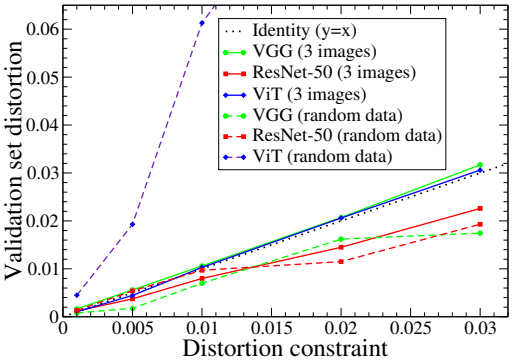
Remarkably, the joint minimizing distribution of the model's weights $p\left(\tilde{\boldsymbol{w}}_{[1:L]}|\boldsymbol{w}_{[1:L]}\right)$ is also spherical since any partitioning of spherical distribution (naturally occurring by the model's layers) remains spherical (Fang et al., 2018, Theorem 2.6). Accordingly, the distortion at the model's output, reflected as angle deviation of the output by eq. (1), is also indifferent to rotations of the weights $\mathbf{w}_{[1:L]}$. Accordingly, the RIQ approach is optimal, as it strives for bin widths $\Delta_{[1:L]}$ and quantization errors that are indifferent to both layer-wise and model-wise rotations. Finally, since the discrete space considered in RIQ approaches the continuous space of the surrogate model in the high-rate regime, then, RIQ is asymptotically optimal.

5 Empirical results

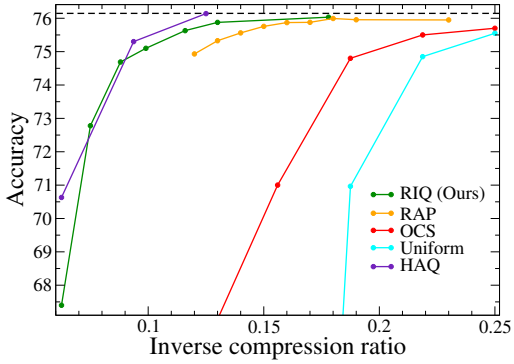
In this section, we evaluate the compression ratio and model accuracy of RIQ with ANS and compare them to relevant baseline results. In all experiments, we use pre-trained models for their relevant tasks. Specifically, for classification tasks we use VGG, Simonyan & Zisserman (2014), ResNet-50, He et al. (2016), and ViT, Dosovitskiy et al. (2020) models, from the torchvision library[1] on the ImageNet data-set (I1k). For detection tasks, we use YOLOv5, Jocher et al. (2022), from NeuralMagic library [2] on the COCO dataset (Lin et al., 2014). For NLP task, we use the DistilBERT

[1] `https://pytorch.org/vision/stable/models.html`

[2] `https://sparsezoo.neuralmagic.com`



(a) Cosine distance of the validation dataset as a function of the distortion constraint (on the calibration dataset). Models include VGG (green circles), ResNet-50 (red squares), and ViT (blue diamonds).



(b) Accuracy vs Compression for ResNet-50 model. RIQ (green) vs. RAP (orange), OCS (red), linear quantization (cyan), and HAQ (purple). HAQ, however, requires training after the quantization.

model, Sanh et al. (2019), on SQuAD dataset, (Rajpurkar et al., 2016). Following RIQ, each quantized layer is encoded with the ANS encoder that achieves asymptotically the entropy limit. An efficient implementation of ANS on GPU [3] was demonstrated by Weißenberger & Schmidt (2019), reaching a decoding rate of over 20 GB/s. For reproduction purposes, we provide a Python code of our algorithm as a supplementary material which includes both the quantization phase (RIQ) and compression phase (ANS). Additional results are given in Appendix A.6.

To measure the resulting distortion by eq. (1) on a validation set as a function of a distortion requirement D, we use two types of calibration data: (a) three real images, sampled from the training, and (b) randomly generated data that follows the Gaussian distribution. In Figure 1a we present the distortion measurements on three models: ResNet-50, VGG, and ViT, where the identity line (black-dotted) is given for reference. As we see, even a small calibration set of three images (solid lines) is sufficient to predict the distortion on the validation set. Further, we see that the randomly generated data may not predict well the resulting distortion on the validation set, leading to either a less compressed model (ResNet-50 and VGG) or a higher distortion (ViT).

In Figure 1b, we evaluate the effect of RIQ on the accuracy and the inverse compression ratio (i.e., the reciprocal of eq. (2)) for a pre-trained ResNet-50 model. Interestingly, the rate-distortion curve reflects well the accuracy-compression tradeoff. For comparison, we depict the accuracy-compression results of the *Relaxed Advanced Pipeline* (RAP) method, Hubara et al. (2021), *Outlier Channel Splitting* (OCS), Zhao et al. (2019), and the *Hardware-aware Automated Quantization* (HAQ), Wang et al. (2019), which requires further training for fine-tuning. Indeed, RIQ provides superior results compared to post-training quantization methods, yet, falls short compared to HAQ. This is since retraining the quantized model creates a different model with a different rate-distortion curve, which is out of the scope of this paper.

In Table 1, we compare RIQ to other relevant baseline methods on the VGG-16 and ResNet-50 models. In this table, we optimized the dual problem of the rate-distortion, i.e., minimizing the distortion (accuracy drop) for a given rate requirement (compression). Thus, the RIQ in Algorithm 1 was slightly modified to obtain a stop condition based on the compression ratio rather than a distortion constraint, and then we measured the resulting accuracy of the quantized models for the VGG and ResNet-50. Noticeably, RIQ outperforms the baseline techniques, yielding better accuracy per compression ratio.

Typical compression ratio and score degradation achieved by RIQ are presented in Table 2 for a variety of models and tasks. Note that the constraint on the quantization output, in the form of a cosine distance in eq. (1) does not translate identically to the drop in each score. While the first is a general distortion metric, the latter is a task-specific measurement. Yet, in general, the scores improve monotonically as the cosine distance decreases. To further assess the potential of RIQ, we evaluate our method on sparse models taken from the Neural-Magic[2]. Notably, the resulting compression of sparse models is significantly higher with relatively small degradation in accuracy.

[3] https://github.com/weissenberger/multians

Table 1: Comparison of Top-1 accuracy on the ImageNet dataset for various quantization techniques.

Model	Comp.	Method	Quant. Acc. (%)	Ref. Acc. (%)	Acc. Drop (%)
VGG-16	32/3	GPFQ Zhang et al. (2022)	70.24	71.59	1.35
		RIQ (Ours)	**71.58**	**71.59**	**0.01**
	32/4	MSE Banner et al. (2019)	70.50	71.60	1.10
		OMSE Choukroun et al. (2019)	71.48	73.48	2.00
		GPFQ Zhang et al. (2022)	70.90	71.59	0.69
		RIQ (Ours)	**71.55**	**71.59**	**0.04**
	32/5	GPFQ Zhang et al. (2022)	71.05	71.59	0.54
		RIQ (Ours)	**71.58**	**71.59**	**0.01**
ResNet-50	32/3	GPFQ Zhang et al. (2022)	70.63	76.13	5.50
		RIQ (Ours)	**74.76**	**76.14**	**1.38**
	32/4	MSE Banner et al. (2019)	73.80	76.10	2.30
		OMSE Choukroun et al. (2019)	73.39	76.01	2.62
		AdaRound Nagel et al. (2020)	75.23	76.07	0.84
		S-AdaQuant Hubara et al. (2021)	75.10	77.20	2.10
		BRECQ Li et al. (2021)	76.29	77.00	0.71
		GPFQ Zhang et al. (2022)	74.35	76.13	1.78
		RIQ (Ours)	**75.61**	**76.14**	**0.53**
	32/5	GPFQ Zhang et al. (2022)	75.26	76.13	0.87
		RIQ (Ours)	**75.95**	**76.14**	**0.19**

Table 2: Compression ratio and accuracy drop achieved by RIQ. Models denoted by asterisk (*) were pruned during training, before quantization.

Model / Dataset	Metric	Distortion constraint	Compression ratio	Quant. Score	FP32 Ref Score	Score Drop
VGG / I1k	Top-1 Acc (%)	0.5%	$\times\mathbf{19.4}$	71.3	71.59	**0.29**
ResNet-50 / I1k	Top-1 Acc (%)	0.5%	$\times\mathbf{7.31}$	75.88	76.14	**0.26**
ViT / I1k	Top-1 Acc (%)	0.5%	$\times\mathbf{6.98}$	81.0	81.07	**0.07**
YOLO / COCO	mAP@.5	0.3%	$\times\mathbf{8.34}$	54.7	55.7	**1.0**
DistilBERT / SQuAD	F1	0.025%	$\times\mathbf{7.96}$	85.0	85.08	**0.08**
VGG* / I1k	Top-1 Acc (%)	0.5%	$\times\mathbf{52.9}$	69.34	69.73	**0.39**
ResNet-50* / I1k	Top-1 Acc (%)	0.5%	$\times\mathbf{41.5}$	75.72	76.14	**0.42**
YOLO* / COCO	mAP@.5	0.3%	$\times\mathbf{16.48}$	52.6	53.5	**0.9**
DistilBERT* / SQuAD	F1	0.025%	$\times\mathbf{19.4}$	84.70	84.92	**0.22**

6 CONCLUSION

In this paper, we have investigated a post-training quantization method that strives to minimize the model's entropy subject to a distortion constraint. The rate-distortion curve was thoroughly analyzed for a surrogate model, showing that the minimizing distribution is a single-letter (scalar) product distribution, constructed as the product of the layer's rotation-invariant distribution, regardless of the number of layers in the model or their distribution. Accordingly, a *rotation-invariant quantization* scheme (RIQ) was introduced, which quantizes each layer in proportion to the layer's norm. That is, layers with a larger norm get lower rate than layers with a small norm. To find the single-letter solution efficiently, we further derived upper and lower bounds of the search space.

Experiments conducted with RIQ show minor accuracy degradation for both computer vision and NLP tasks. When applying RIQ on sparse models, the compression ratio reached remarkable values (up to $\times 52.9$) with a score degradation of less than 1%. These superb compression rates together with the negligible accuracy drop, are attractive and practical. In particular, RIQ minimizes the entropy of each layer, and hence is suitable for per-layer execution of compressed models, thus, facilitating the deployment of large models on lean servers or edge devices with limited memory.

REFERENCES

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems*, 32, 2019.

Chaim Baskin, Brian Chmiel, Evgenii Zheltonozhskii, Ron Banner, Alex M. Bronstein, and Avi Mendelson. Cat: Compression-aware training for bandwidth reduction, 2019. URL https://arxiv.org/abs/1909.11481.

Chaim Baskin, Natan Liss, Eli Schwartz, Evgenii Zheltonozhskii, Raja Giryes, Alex M Bronstein, and Avi Mendelson. Uniq: Uniform noise injection for non-uniform quantization of neural networks. *ACM Transactions on Computer Systems (TOCS)*, 37(1-4):1–15, 2021.

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Gabriel Caffarena, Carlos Carreras, Juan A López, and Ángel Fernández. Sqnr estimation of fixed-point dsp algorithms. *EURASIP Journal on Advances in Signal Processing*, 2010:1–12, 2010.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework, 2020. URL https://arxiv.org/abs/2001.00281.

Michael Alan Chang, Domenic Bottini, Lisa Jian, Pranay Kumar, Aurojit Panda, and Scott Shenker. How to train your dnn: The network operator edition. *arXiv preprint arXiv:2004.10275*, 2020.

Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3009–3018. IEEE, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

Sam Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. In *NeurIPS*, 2021a.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 33:9851–9864, 2020.

Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Multi-objective bayesian optimization over high-dimensional search spaces. *arXiv preprint arXiv:2109.10964*, 2021b.

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

Jarek Duda. Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2013.

David Eriksson and Martin Jankowiak. High-dimensional bayesian optimization with sparse axis-aligned subspaces. In *UAI*, 2021.

Angela Fan, Pierre Stock, Benjamin Graham, Edouard Grave, Rémi Gribonval, Hervé Jégou, and Armand Joulin. Training with quantization noise for extreme model compression. *CoRR*, abs/2004.07320, 2020. URL https://arxiv.org/abs/2004.07320.

Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC, 2018.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

David Freedman and Persi Diaconis. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476, 1981.

Weihao Gao, Yu-Han Liu, Chong Wang, and Sewoong Oh. Rate distortion for model compression: From theory to practice. In *International Conference on Machine Learning*, pp. 2102–2111. PMLR, 2019.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2015. URL https://arxiv.org/abs/1510.00149.

Babak Hassibi, David G Stork, and Gregory J Wolff. Optimal brain surgeon and general network pruning. In *IEEE international conference on neural networks*, pp. 293–299. IEEE, 1993.

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

Stephen Montgomery-Smith (https://stephenmontgomerysmith.github.io/). Finding the rotation matrix in n-dimensions. Mathematics Stack Exchange, 2016. URL https://math.stackexchange.com/q/598782. URL:https://math.stackexchange.com/q/598782 (version: 2016-06-19).

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Accurate post training quantization with small calibration sets. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4466–4475. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hubara21a.html.

Yerlan Idelbayev and Miguel A Carreira-Perpinán. Low-rank compression of neural nets: Learning the rank of each layer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8049–8059, 2020.

Yerlan Idelbayev, Pavlo Molchanov, Maying Shen, Hongxu Yin, Miguel A Carreira-Perpinán, and Jose M Alvarez. Optimal quantization using scaled codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12095–12104, 2021.

Berivan Isik, Tsachy Weissman, and Albert No. An information-theoretic justification for model pruning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3821–3846. PMLR, 2022.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, February 2022. URL https://doi.org/10.5281/zenodo.6222936.

VN Koshelev. Quantization with minimal entropy. *Probl. Pered. Inform*, 14:151–156, 1963.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *ArXiv*, abs/1806.08342, 2018.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Dongsoo Lee, Se Jung Kwon, Byeongwook Kim, and Gu-Yeon Wei. Learning low-rank approximation for CNNs. *arXiv preprint arXiv:1905.10145*, 2019.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Shaohui Lin, Rongrong Ji, Chao Chen, Dacheng Tao, and Jiebo Luo. Holistic cnn compression via low-rank decomposition with knowledge transfer. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2889–2905, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Devira Anggi Maharani, Carmadi Machbub, Pranoto Hidaya Rusmin, and Lenni Yulianti. Improving the capability of real-time face masked recognition using cosine distance. *2020 6th International Conference on Interactive Digital Media (ICIDM)*, pp. 1–6, 2020.

Springer Melvin Dale. *The algebra of random variables*. Wiley, 1979.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pp. 7197–7206. PMLR, 2020.

NVIDIA. Pytorch quantization - functionalities. docs.nvidia.com, 2021. URL `https://docs.nvidia.com/deeplearning/tensorrt/pytorch-quantization-toolkit/docs/userguide.html`.

Deniz Oktay, Johannes Ballé, Saurabh Singh, and Abhinav Shrivastava. Scalable model compression by entropy penalized reparameterization. *arXiv preprint arXiv:1906.06624*, 2019.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Knowledge distillation beyond model compression. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6136–6143. IEEE, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Wendi Usino, Anton Satria Prabuwono, Khalid Hamed S. Allehaibi, Arif Bramantoro, A. Hasniaty, and Wahyu Amaldi. Document similarity detection using k-means and cosine distance. *International Journal of Advanced Computer Science and Applications*, 2019.

Devesh Walawalkar, Zhiqiang Shen, and Marios Savvides. Online ensemble model compression using knowledge distillation. In *European Conference on Computer Vision*, pp. 18–35. Springer, 2020.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8612–8620, 2019.

André Weißenberger and Bertil Schmidt. Massively parallel ans decoding on gpus. In *Proceedings of the 48th International Conference on Parallel Processing*, pp. 1–10, 2019.

Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. Integer quantization for deep learning inference: Principles and empirical evaluation. *arXiv preprint arXiv:2004.09602*, 2020.

Jinjie Zhang, Yixuan Zhou, and Rayan Saab. Post-training quantization for neural networks with provable guarantees. *arXiv preprint arXiv:2201.11113*, 2022.

Yanfu Zhang, Shangqian Gao, and Heng Huang. Exploration and estimation for model compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 487–496, 2021.

Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *International conference on machine learning*, pp. 7543–7552. PMLR, 2019.

A APPENDIX

In this section, we provide rigorous proofs for the theorems and the statements herein. Further, we present additional results for RIQ.

A.1 PROOF OF LEMMA 1

Proof. Let $\mathbf{w}_\ell$ be the realization of the weights vector of layer ℓ, and $\hat{\mathbf{w}}_\ell$ be the quantized representation of those weights, where θ_ℓ denotes the angle between those vectors. Before diving into the cosine distance analysis, let us revisit the mean squared error analysis of the uniform quantizer in Section 3.3, and extend it to the multivariate case. In this case, the distortion is

$$\begin{aligned}\|\mathbf{w}_\ell - \hat{\mathbf{w}}_\ell\|^2 &= n_\ell \cdot \frac{1}{n_\ell}\sum_{i=1}^{n_\ell} |\mathrm{w}_{\ell,i} - \hat{\mathrm{w}}_{\ell,i}|^2 \\ &\overset{(a)}{=} n_\ell \cdot \mathbb{E}\,|\mathrm{w}_{\ell,j} - \hat{\mathrm{w}}_{\ell,j}|^2 + o(1) \\ &\overset{(b)}{=} n_\ell \cdot \Delta_\ell^2/12 + o(1) \end{aligned} \tag{11}$$

where (a) follows from the law of large numbers, and (b) follows by the analysis of the scalar uniform quantizer, given in (Polyanskiy & Wu, 2014, Ch. 25.1).

This observation is useful for analyzing the cosine distance between $\mathbf{w}_\ell$ and $\hat{\mathbf{w}}_\ell$. Specifically, since

$$\|\mathbf{w}_\ell - \hat{\mathbf{w}}_\ell\|^2 = \|\mathbf{w}_\ell\|^2 + \|\hat{\mathbf{w}}_\ell\|^2 - 2\|\mathbf{w}_\ell\| \cdot \|\hat{\mathbf{w}}_\ell\| \cos(\theta_\ell),$$

assuming $\|\hat{\mathbf{w}}_\ell\| = \|\mathbf{w}_\ell\| + o(\|\mathbf{w}_\ell\|)$, yields

$$\begin{aligned}\|\mathbf{w}_\ell - \hat{\mathbf{w}}_\ell\|^2 &= 2\|\mathbf{w}_\ell\|^2 + o(\|\mathbf{w}_\ell\|^2) - 2\|\mathbf{w}_\ell\|^2 \cos(\theta_\ell) + o(\|\mathbf{w}_\ell\|^2) \\ &= 2\|\mathbf{w}_\ell\|^2 \cdot (1 - \cos(\theta_\ell)) + o(\|\mathbf{w}_\ell\|^2).\end{aligned}$$

Hence, normalizing both sides by $2\|\mathbf{w}_\ell\|^2$, we obtain that

$$(1 - \cos(\theta_\ell)) = \frac{\|\mathbf{w}_\ell - \hat{\mathbf{w}}_\ell\|^2}{2\|\mathbf{w}_\ell\|^2} + o(1). \tag{12}$$

Combining the analysis of eq. (11) with eq. (12), we obtain

$$(1 - \cos(\theta_\ell)) = \frac{\|\mathbf{w}_\ell - \hat{\mathbf{w}}_\ell\|^2}{2\|\mathbf{w}_\ell\|^2} + o(1) = \frac{\Delta_\ell^2 \cdot n_\ell}{24 \cdot \|\mathbf{w}_\ell\|^2} + o(1).$$

By denoting $\epsilon_\ell = 1 - \cos(\theta_\ell)$, and omitting the little order $o(1)$, the lemma follows. □

A.2 PROOF OF PROPOSITION 1

Proof. The layer whose quantization error converges last to ϵ_0 dictates when to stop the search. Specifically, when k is sufficiently large in eq. (7), the error in layer ℓ^* reaches $\sqrt{\epsilon_{\ell^*}} = o(\epsilon_0) + \epsilon_0$. That is where ϵ_0 becomes dominant. At this point, we say that the error has converged for all layers (as it converged even at the largest layer ℓ^*). Since $\epsilon_0 \leq 1$ in the cosine distance criterion, we choose the little order to be $o(\epsilon_0) = \epsilon_0 \cdot \sqrt{\epsilon_0}$, and hence, k can be bounded from above by

$$\frac{1}{k}\sqrt{n_{\ell^*}/24} + \epsilon_0 \geq o(\epsilon_0) + \epsilon_0,$$

which happens when $k \leq \sqrt{n_{\ell^*}/24}/(\epsilon_0 \cdot \sqrt{\epsilon_0})$.

In our experiments, we let $\epsilon_0 = 0.01$, hence, the upper limit is simply $k \leq 1000 \cdot \sqrt{n_{\ell^*}/24}$ [4]

For a lower bound, we use again the fact that $\epsilon_\ell \leq 1$. Thus, focusing on layer ℓ^*, we observe that

$$\frac{1}{k} \cdot \sqrt{n_{\ell^*}/24} + \epsilon_0 \leq 1,$$

which happens as long as $k \geq \sqrt{n_{\ell^*}/24}/(1 - \epsilon_0)$. This completes the proof. □

[4] When the original weights are represented with R bit symbols, then, choosing $\epsilon_0 = 0$, yields a trivial upper bound $k_{\max}$, which is the largest number that can be represented with R bits, e.g., $\lceil k_{\max} \rfloor \leq 2^{31}$, when using 32 bits integer.

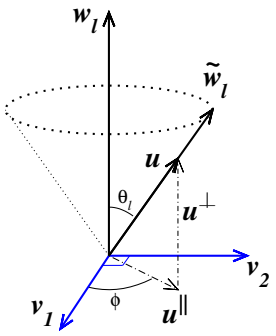


Figure 2: Illustration of the projection of $\tilde{\mathbf{w}}_\ell$ onto the arbitrary perpendicular vectors $\mathbf{v}_1$ and $\mathbf{v}_2$.

A.3 Proof of Proposition 2

Proof. By (Fang et al., 2018, Theorem 4.3), a necessary and sufficient condition for $\tilde{\mathbf{w}}_\ell \big| \mathbf{w}_\ell$ to have a spherical rotation-invariant distribution is when

$$p\left(\langle \tilde{\mathbf{w}}_\ell, \mathbf{v}_1\rangle \middle| \langle \tilde{\mathbf{w}}_\ell, \mathbf{v}_2\rangle, \mathbf{w}_\ell\right) \overset{d}{=} p\left(-\langle \tilde{\mathbf{w}}_\ell, \mathbf{v}_1\rangle \middle| \langle \tilde{\mathbf{w}}_\ell, \mathbf{v}_2\rangle, \mathbf{w}_\ell\right),$$

for any pair of perpendicular vectors $\mathbf{v}_1 \neq 0$ and $\mathbf{v}_2 \neq 0$.

Consider the model in eq. (9), any orthogonal transformation $\mathbf{U}(\theta_\ell \big| \mathbf{w}_\ell)$ can be represented by an orthonormal basis, obtained by the Gram-Schmidt process. That is, finding two orthonormal vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ that span the plane of rotation generated by $\mathbf{w}_\ell$ and some $\mathbf{w}'_\ell$ that is θ_ℓ away from $\mathbf{w}_\ell$, and then, extend this basis to $\mathbb{R}^{n_\ell}$. This allows us to consider the rotation in the plane generated by those vectors, with respect to the extended basis (https://stephenmontgomerysmith.github.io/). Accordingly, let $\mathbf{u}_1 = \frac{\mathbf{w}_\ell}{\|\mathbf{w}_\ell\|}$ and $\mathbf{u}_2 = \frac{\mathbf{w}'_\ell - \langle \mathbf{u}_1, \mathbf{w}'_\ell\rangle \mathbf{u}_1}{\|\mathbf{w}'_\ell - \langle \mathbf{u}_1, \mathbf{w}'_\ell\rangle \mathbf{u}_1\|}$, then

$$\mathbf{U}(\theta_\ell \big| \mathbf{w}_\ell) = \mathbf{I}_{n_\ell} - \mathbf{u}_1\mathbf{u}_1^T - \mathbf{u}_2\mathbf{u}_2^T + [\mathbf{u}_1, \mathbf{u}_2]\, \mathbf{R}_{\theta_\ell}\, [\mathbf{u}_1, \mathbf{u}_2]^T, \tag{13}$$

where $\mathbf{I}_{n_\ell}$ is the $n_\ell \times n_\ell$ identity matrix and $\mathbf{R}_{\theta_\ell}$ is the rotation matrix

$$\mathbf{R}_{\theta_\ell} = \begin{bmatrix} \cos(\theta_\ell) & -\sin(\theta_\ell) \\ \sin(\theta_\ell) & \cos(\theta_\ell) \end{bmatrix},$$

that rotates at a scalar angle θ_ℓ, and $[\mathbf{u}_1, \mathbf{u}_2]$ is $n_\ell \times 2$ matrix whose columns are $\mathbf{u}_1$ and $\mathbf{u}_2$, respectively. Plugging eq. (13) to eq. (9), and noting that $\frac{\mathbf{w}_\ell}{\|\mathbf{w}_\ell\|} = \mathbf{u}_1$, we obtain

$$\tilde{\mathbf{w}}_\ell = \|\hat{\mathbf{w}}_\ell\| \cdot (\cos(\theta_\ell)\mathbf{u}_1 + \sin(\theta_\ell)\mathbf{u}_2) \tag{14}$$

To simplify notation, let $\mathbf{u} \triangleq (\cos(\theta_\ell)\mathbf{u}_1 + \sin(\theta_\ell)\mathbf{u}_2)$, and note that for any perpendicular pair $\mathbf{v}_1, \mathbf{v}_2$, the vector $\mathbf{u}$ can be decomposed to $\mathbf{u} = \mathbf{u}^{\parallel} + \mathbf{u}^{\perp}$, where $\mathbf{u}^{\parallel}$ resides in the plane generated by $\mathbf{v}_1$ and $\mathbf{v}_2$, and $\mathbf{u}^{\perp}$ resides in the null-space of this plane. For illustration, see Figure 2. Hence,

$$\tilde{\mathbf{w}}_\ell = \|\hat{\mathbf{w}}_\ell\| \cdot (\mathbf{u}^{\parallel} + \mathbf{u}^{\perp}) \tag{15}$$

Accordingly, we have

$$\begin{aligned}
& p\left(\langle \tilde{\mathbf{w}}_\ell, \mathbf{v}_1\rangle \middle| \langle \tilde{\mathbf{w}}_\ell, \mathbf{v}_2\rangle, \mathbf{w}_\ell\right) \qquad (16)\\
&\overset{(a)}{=} p\left(\langle \|\hat{\mathbf{w}}_\ell\| \cdot (\mathbf{u}^{\parallel} + \mathbf{u}^{\perp}), \mathbf{v}_1\rangle \middle| \langle \|\hat{\mathbf{w}}_\ell\| \cdot (\mathbf{u}^{\parallel} + \mathbf{u}^{\perp}), \mathbf{v}_2\rangle, \mathbf{w}_\ell\right)\\
&\overset{(b)}{=} p\left(\langle \|\hat{\mathbf{w}}_\ell\| \cdot \mathbf{u}^{\parallel}, \mathbf{v}_1\rangle \middle| \langle \|\hat{\mathbf{w}}_\ell\| \cdot \mathbf{u}^{\parallel}, \mathbf{v}_2\rangle, \mathbf{w}_\ell\right)\\
&\overset{(c)}{=} p\left(\|\hat{\mathbf{w}}_\ell\| \cdot \|\mathbf{u}^{\parallel}\| \cdot \cos(\phi) \middle| \|\hat{\mathbf{w}}_\ell\| \cdot \|\mathbf{u}^{\parallel}\| \cdot \sin(\phi), \mathbf{w}_\ell\right)\\
&\overset{(d)}{=} p\left(\|\hat{\mathbf{w}}_\ell\| \cdot \|\mathbf{u}^{\parallel}\| \cdot \cos(\pi - \phi) \middle| \|\hat{\mathbf{w}}_\ell\| \cdot \|\mathbf{u}^{\parallel}\| \cdot \sin(\pi - \phi), \mathbf{w}_\ell\right)
\end{aligned}$$

$$\stackrel{(e)}{=} p\left(-\|\hat{\mathbf{w}}_\ell\| \cdot \|\mathbf{u}^{\|}\| \cdot \cos(\phi) \middle| \|\hat{\mathbf{w}}_\ell\| \cdot \|\mathbf{u}^{\|}\| \cdot \sin(\phi), \mathbf{w}_\ell\right)$$
$$= p\left(-\left\langle \hat{\mathbf{w}}_\ell, \mathbf{v}_1 \right\rangle \middle| \left\langle \hat{\mathbf{w}}_\ell, \mathbf{v}_2 \right\rangle, \mathbf{w}_\ell\right) \tag{17}$$

where (a) follows by eq. (15). (b) follows by the linearity of the inner product and since $\mathbf{u}^{\perp}$ is perpendicular to both $\mathbf{v}_1$ and $\mathbf{v}_2$ (c) follows due to the orthogonality of the basis $\mathbf{v}_1$ and $\mathbf{v}_2$, where ϕ is the angle between $\mathbf{u}^{\|}$ and $\mathbf{v}_1$. (d) follows since the angle between $\mathbf{u}^{\|}$ and an arbitrary $\mathbf{v}_1$ is arbitrary, and hence, every angle has the same distribution. (e) follows by trigonometric identities for the cosine and sine function.

Thus, Proposition 2 follows. □

A.4 Proof of Theorem 1

Proof. Assuming the weights of each layer $\mathbf{w}_\ell$ are statistically independent of the weights of the other layers, then, by the properties of the mutual information, we have

$$I\left(\mathbf{w}_{[1:L]}; \tilde{\mathbf{w}}_{[1:L]}\right) = H\left(\mathbf{w}_{[1:L]}\right) - H\left(\mathbf{w}_{[1:L]} \middle| \tilde{\mathbf{w}}_{[1:L]}\right) \tag{18}$$
$$= \sum_{\ell=1}^{L} H\left(\mathbf{w}_\ell\right) - \sum_{\ell=1}^{L} H\left(\mathbf{w}_\ell \middle| \mathbf{w}_{[1:\ell-1]}, \tilde{\mathbf{w}}_{[1:L]}\right) \tag{19}$$
$$\geq \sum_{\ell=1}^{L} H\left(\mathbf{w}_\ell\right) - \sum_{\ell=1}^{L} H\left(\mathbf{w}_\ell \middle| \tilde{\mathbf{w}}_\ell\right) \tag{20}$$
$$= \sum_{\ell=1}^{L} I\left(\mathbf{w}_\ell; \tilde{\mathbf{w}}_\ell\right) \tag{21}$$
$$\geq \sum_{\ell=1}^{L} R(D_\ell) \tag{22}$$

where eq. (20) follows since conditioning reduces entropy. Note, however, that eq. (20) can be attained with equality by letting $p\left(\mathbf{w}_{[1:L]} \middle| \tilde{\mathbf{w}}_{[1:L]}\right) = \prod_{\ell=1}^{L} p\left(\mathbf{w}_\ell \middle| \tilde{\mathbf{w}}_\ell\right)$. Consequently, the minimizing distribution in eq. (10) is a product distribution (Polyanskiy & Wu, 2014, Theorem 5.1 (2)).

Interestingly, eq. (22) implies that the optimal solution has a simple form of a *layer-by-layer solution*, which significantly simplifies the problem at hand. Finding the solution, of course, requires formulating the relation between D and D_ℓ of each layer ℓ, and hence, the resulting rate $R(D_\ell)$ in each layer. Moreover, since each layer obtains a different rate, it implies that the optimal solution is indeed a *mixed-precision solution*, where each layer can be considered independently, and hence, the minimizing distribution is product distribution.

Let $p\left(\tilde{\mathbf{w}}_1 \middle| \mathbf{w}_1\right) \cdot p\left(\tilde{\mathbf{w}}_2 \middle| \mathbf{w}_2\right) \cdots p\left(\tilde{\mathbf{w}}_L \middle| \mathbf{w}_L\right)$ be a distribution that satisfies the distortion requirement D, for which the induced cosine distance in each layer ℓ is at most D_ℓ for some values D_ℓ. When $\frac{1}{L}\sum_{\ell=1}^{L} D_\ell \leq D$, due to the convexity of the cosine distance for $|\theta_\ell| \leq \pi/2$, by Jensen inequality $\left(1 - \cos\left(\frac{1}{L}\sum_{\ell=1}^{L} \theta_\ell\right)\right) \leq \frac{1}{L}\sum_{\ell=1}^{L}\left(1 - \cos(\theta_\ell)\right)$. In words, the average angle also satisfies the distortion constraint D. Interestingly, the cosine distance is rotation-invariant by definition, as the angle between vectors does not change when they are rotated together. This further hints that the minimizing distribution should also be a rotation-invariant distribution as follows.

By eq. (14), given $\mathbf{w}_\ell$ (and hence, $\mathbf{u}_1$), the probability of $\tilde{\mathbf{w}}_\ell$ is determined by the probability of the rotation angle θ_ℓ and the length $\|\hat{\mathbf{w}}_\ell\|$. Specifically, for any vector $\mathbf{s}_\ell \in \mathbb{R}^{n_\ell}$, the density function of this product, if exists, is (Melvin Dale, 1979, Ch. 4.1)

$$p_{\tilde{\mathbf{w}}_\ell|\mathbf{w}_\ell}(\mathbf{s}_\ell) = \int_0^\infty p_{\|\hat{\mathbf{w}}_\ell\||\mathbf{w}_\ell}(h) \cdot p_{(\cos(\theta_\ell)\mathbf{u}_1 + \sin(\theta_\ell)\mathbf{u}_2)|\mathbf{w}_\ell}(\mathbf{s}_\ell/h) \cdot \frac{1}{h}\mathrm{d}h, \tag{23}$$

where the rotation θ_ℓ occurs on $\mathbb{R}^2$, rotating about $(n_\ell - 2)$-dimensional subspace. Further, note that the dimension n_ℓ is dictated only by the given $\mathbf{w}_\ell$. Accordingly, since each layer ℓ resides at a different dimension n_ℓ, it is impossible to consider the average layer distribution directly, as done for the vector case, e.g., as considered in (Polyanskiy & Wu, 2014, Ch. 5). Nevertheless, since

the rotation of θ_ℓ is done on $\mathbb{R}^2$ in each layer ℓ, which is described by the rotation matrix $\mathbf{R}_{\theta_\ell}$ in eq. (13), it is still beneficial to consider the average distribution of θ_ℓ over the layers, to allow a similar treatment to (Polyanskiy & Wu, 2014, Ch. 5), as follows.

To bound the mutual information, the density of $(\cos(\theta_\ell)\mathbf{u}_1 + \sin(\theta_\ell)\mathbf{u}_2)\,|\mathbf{w}_\ell$ should be expressed first in terms of the density of $\cos(\theta_\ell)|\mathbf{w}_\ell$. Examining eq. (14), we note that by the transformation of random variables formula,

$$p_{(\cos(\theta_\ell)\mathbf{u}_1+\sin(\theta_\ell)\mathbf{u}_2)|\mathbf{w}_\ell}\left(\mathbf{s}_\ell/h\right) = p_{\cos(\theta_\ell)|\mathbf{w}_\ell}\left(\mathbf{u}_1^T\mathbf{s}_\ell/h\right). \tag{24}$$

Considering the high rate regime, where Δ_ℓ is sufficiently small, and thus, $\|\hat{\mathbf{w}}_\ell\| = \|\mathbf{w}_\ell\| + o(1)$, then, the density function in eq. (23) becomes

$$\begin{aligned}
p_{\tilde{\mathbf{w}}_\ell\left|\mathbf{w}_\ell\right.}(\mathbf{s}_\ell) &= p_{\|\hat{\mathbf{w}}_\ell\|\cdot(\cos(\theta_\ell)\mathbf{u}_1+\sin(\theta_\ell)\mathbf{u}_2)|\mathbf{w}_\ell}(\mathbf{s}_\ell) && (25)\\
&= \int_0^\infty p_{\|\hat{\mathbf{w}}_\ell\||\mathbf{w}_\ell}\left(h\right)\cdot p_{(\cos(\theta_\ell)\mathbf{u}_1+\sin(\theta_\ell)\mathbf{u}_2)|\mathbf{w}_\ell}\left(\mathbf{s}_\ell/h\right)\cdot\frac{1}{h}\mathrm{d}h && (26)\\
&\overset{(a)}{=} \int_0^\infty p_{\|\hat{\mathbf{w}}_\ell\|\|\mathbf{w}_\ell\|}\left(h\right)\cdot p_{(\cos(\theta_\ell)\mathbf{u}_1+\sin(\theta_\ell)\mathbf{u}_2)|\mathbf{w}_\ell}\left(\mathbf{s}_\ell/h\right)\cdot\frac{1}{h}\mathrm{d}h && (27)\\
&\overset{(b)}{\approx} \int_0^\infty \delta\left(h-\|\mathbf{w}_\ell\|\right)\cdot p_{(\cos(\theta_\ell)\mathbf{u}_1+\sin(\theta_\ell)\mathbf{u}_2)|\mathbf{w}_\ell}\left(\mathbf{s}_\ell/h\right)\cdot\frac{1}{h}\mathrm{d}h && (28)\\
&\overset{(c)}{=} \int_0^\infty \delta\left(h-\|\mathbf{w}_\ell\|\right)\cdot p_{\cos(\theta_\ell)|\mathbf{w}_\ell}\left(\mathbf{u}_1^T\mathbf{s}_\ell/h\right)\cdot\frac{1}{h}\mathrm{d}h && (29)\\
&\overset{(d)}{=} p_{\cos(\theta_\ell)|\mathbf{w}_\ell}\left(\mathbf{u}_1^T\mathbf{s}_\ell/\|\mathbf{w}_\ell\|\right)\cdot\|\mathbf{w}_\ell\|^{-1} && (30)
\end{aligned}$$

where (a) follows since the norm $\|\mathbf{w}_\ell\|$ is a function of the given $\mathbf{w}_\ell$. (b) follows since the uncertainty about $\|\hat{\mathbf{w}}_\ell\|$ given $\|\mathbf{w}_\ell\|$ is negligible, and hence, $p_{\|\hat{\mathbf{w}}_\ell\|\|\mathbf{w}_\ell\|}(h) \approx \delta(h-\|\mathbf{w}_\ell\|)$, i.e., the conditional density is approximately the Dirac delta function. (c) follows by eq. (24). (d) follows by the characteristics of the Dirac delta function.

Hence, averaging $p\left(\tilde{\mathbf{w}}_\ell\middle|\mathbf{w}_\ell\right)$ over the layers can be approximated by averaging the rotations in $\mathbb{R}^2$ over the model. Accordingly, let

$$\bar{p}_{\tilde{\mathbf{w}}_{[1:L]}\left|\mathbf{w}_{[1:L]}\right.}(\mathbf{s}_{[1:L]}) \triangleq \frac{1}{L}\sum_{\ell=1}^{L} p_{\cos(\theta_\ell)|\mathbf{w}_\ell}\left(\frac{\mathbf{w}_\ell^T\mathbf{s}_\ell}{\|\mathbf{w}_\ell\|^2}\right)\cdot\|\mathbf{w}_\ell\|^{-1}. \tag{31}$$

Then, by the convexity of the rate-distortion function (Cover & Thomas, 2006, Theorem 2.7.4), $\bar{p}\left(\tilde{\mathbf{w}}_{[1:L]}\middle|\mathbf{w}_{[1:L]}\right)$ can only reduce the mutual information in eq. (21). Specifically,

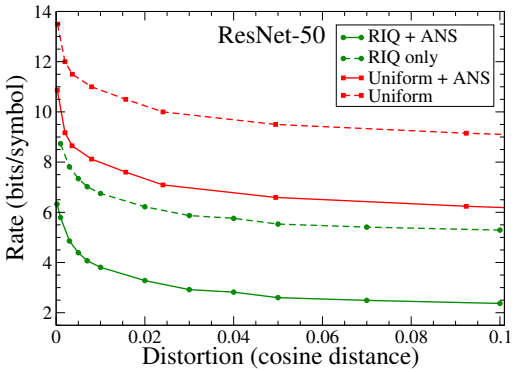
$$\sum_{\ell=1}^{L} I_{p\left(\tilde{\mathbf{w}}_\ell\middle|\mathbf{w}_\ell\right)}\left(\mathbf{w}_\ell;\tilde{\mathbf{w}}_\ell\right) \geq L\cdot I_{\bar{p}\left(\tilde{\mathbf{w}}_{[1:L]}\middle|\mathbf{w}_{[1:L]}\right)}\left(\mathbf{w}_{[1:L]};\tilde{\mathbf{w}}_{[1:L]}\right)$$

where $I_p(\cdot;\cdot)$ denotes explicitly the mutual information under probability p. Thus, the *infimum rate has a form of a scalar (single-letter) rate*.

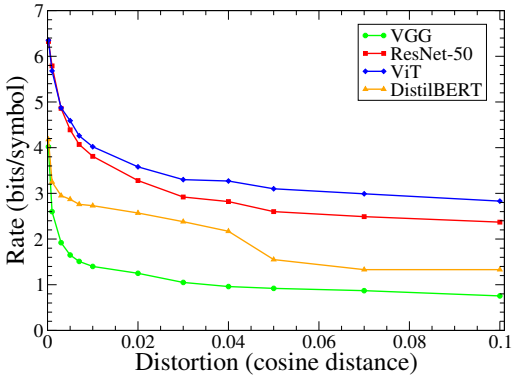
Moreover, since averaging over more rotations should further reduce the mutual information by its convexity, then, the minimizing $p\left(\tilde{\mathbf{w}}_{[1:L]}\middle|\mathbf{w}_{[1:L]}\right)$ can be chosen to be rotation-invariant (Polyanskiy & Wu, 2014, Ch. 5.2). Consequently, the unique minimizing distribution $p\left(\tilde{\mathbf{w}}_{[1:L]}\right)$ is also rotation-invariant. Remarkably, (Fang et al., 2018, Theorem 2.6) states that when partitioning a spherical rotation-invariant distribution (naturally, according to the layers $\tilde{\mathbf{w}}_\ell$), then its components also have a spherical rotation-invariant distribution. This coincides with Proposition 2, which proves that the partitioning satisfies this property.

Accordingly, the unique minimizing distribution $p(\tilde{\mathbf{w}}_{[1:L]}\big|\mathbf{w}_{[1:L]})$ of the rate-distortion function is a product distribution over the layers, where each term ℓ is a spherical rotation-invariant distribution. This completes the proof.

□



(a) Rate-distortion curve for ResNet-50 model obtained for RIQ (green circles) as well as Uniform linear quantization (red squares). Rates are presented for both the quantized model (dashed) as well as following an ANS compression.



(b) Rate distortion curves obtained by RIQ + ANS, for a variety of models: VGG (green circles), ResNet-50 (red squares), ViT (blue diamonds), and DistilBERT (orange triangles).

A.5 Relation to Other Error Criteria

Remark. *The proof of Lemma 1 in Appendix A.1 may serve as a proxy to other error criteria such as the* Signal to Quantization Noise Ratio *(SQNR), Caffarena et al. (2010). Specifically, similar to the proof of Lemma 1, the resulting connection between the scale Δ_ℓ and the SQNR ϵ'_ℓ in each layer ℓ is*

$$\epsilon'_\ell \triangleq \frac{\|\boldsymbol{w}_\ell - \hat{\boldsymbol{w}}_\ell\|}{\|\boldsymbol{w}_\ell\|} = \sqrt{\frac{\Delta_\ell^2}{12} \cdot \frac{n_\ell}{\|\boldsymbol{w}_\ell\|^2}}$$

Or, equivalently,

$$\Delta_\ell = \epsilon'_\ell \|\boldsymbol{w}_\ell\| \sqrt{12/n_\ell}$$

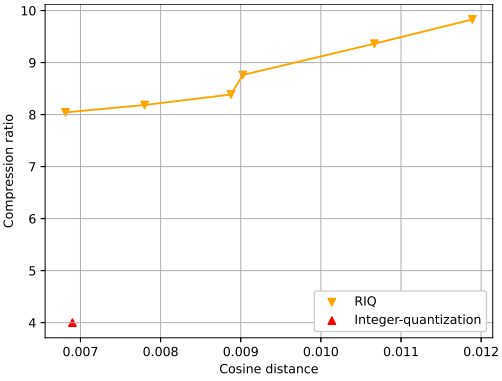
A.6 Additional Results

A.6.1 Decomposing the Rate-Distortion Curve

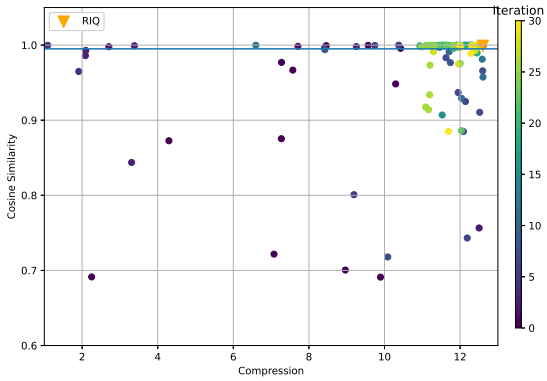
The key steps of lossy compression are quantization and compression. In the quantization phase, the RIQ approach is minimizing the overall model's entropy by allocating a small number of unique symbols for large-norm layers. To achieve (asymptotically) this entropy limit, we utilize the ANS (lossless) entropy encoder. In this section, we evaluate the contribution of each step to the rate-distortion tradeoff. Namely, the average rate per (quantized) symbol before and after ANS. At run-time, when a certain layer is required, it is decoded and represented at a rate according to RIQ. If this rate is below 8 bits/symbol, it enables significant acceleration by performing 8 bits integer operations, as discussed in Appendix A.6.2.

Figure 3a depicts the rate-distortion curve for ResNet-50, decomposed to the quantization step (dashed lines) and the resulting compression step, following the quantization step (solid lines). As a baseline, the uniform scalar quantization (red color) is given for comparison with RIQ (green color). Interestingly, RIQ (dashed green line) outperforms the uniform quantization (dashed red line) by about ~ 4 bits/symbol and even its resulting compressed size by about ~ 1 bit/symbol. Indeed, the latter indicates that uniform quantization does not minimize the model's entropy. Applying the ANS compression following RIQ reduces additional ~ 3 bits/symbol (solid green line), which according to our analysis is the minimum entropy possible for a given distortion. Moreover, our method achieves a reduction of about ~ 8 bits/symbol compared to uniform scalar quantization alone, and an additional ~ 4 bits/symbol when ANS is applied to the uniformly quantized weights.

The rate-distortion curves for various models, in particular, the VGG (green circles), ResNet-50 (red squares), ViT (blue diamonds), and DistilBERT (orange triangles) are given in Figure 3b. As expected, the curves decrease monotonously, reaching an impressive compression rate of less than 8 bits/symbol on average even for extremely low cosine distance in all presented models.



(a) The compression ratio as a function of cosine distance. The left-bottom red triangle depicts the resulting distance of 0.0069 achieved by the baseline with a compression ratio of $\times 4$. The orange upside-down triangles depict the cosine distance and compression ratio attained by RIQ with the ANS compression. The orange line depicts the trend line.



(b) MOBO optimization process. Interestingly, MOBO converges at the few last iterations to $\times 12$ compression, with a highest value of $\times 12.61$. On the other hand, RIQ reaches practically the same compression ratio in a few seconds.

A.6.2 Rotation Invariant Quantization with Quantized Activations

Quantizing both the NN model's weights and its activations can further accelerate the inference, Wu et al. (2020); Nagel et al. (2020); Krishnamoorthi (2018). Nevertheless, in this case, the quantization error of both the weights and the activation affects the model's output. In the seminal work of Wu et al. (2020), the authors utilized the KL distance for quantizing the activations to minimize the information loss at the output. In this section, we examine the RIQ approach, combining it with activation quantization.

To demonstrate, we use the NVIDIA (2021) quantization library for the ResNet-50 model with a "mini ImageNet" validation set, which comprises one image per class and a total of 1000 images. We evaluate this library's performance as a baseline, where the activations are quantized by the KL-distance criterion, and the weights are quantized to 8-bit linearly. The resulting cosine distance at the output of this baseline is 0.69%. For comparison, this reference value is given as the distortion requirement to RIQ. In particular, to integrate RIQ, the activations are quantized as the baseline, and then, we run RIQ according to Algorithm 1. This way, RIQ is aware of the activations' quantization error during its search for the single-letter solution. Note that to facilitate the acceleration of int8 operations, RIQ must yield a quantization rate of up to 8 bits/symbol. In case a certain layer requires a higher rate, we simply perform linear uniform quantization to 8 bits (without clipping), as the baseline does.

Figure 4a characterizes the compression ratio as a function of cosine distance. The leftmost point reflects a cosine distance of 0.0069 achieved by the baseline of NVIDIA (2021). Remarkably, the RIQ attains superior compression with relatively low distortion even when the activations are quantized. In run-time, of course, the reconstructed values are represented again by 8-bit value, and hence, the significant acceleration of Wu et al. (2020) is still valid.

A.6.3 Comparison with Multi-Objective Bayesian Optimization

In this section, we utilize the *Multi-Objective Bayesian Optimization* (MOBO) tool, described in Daulton et al. (2020) to compress NN models, and compare results with RIQ. To compress models with MOBO, we set two objective functions for it. The first objective is minimizing the cosine distance in eq. (1). The second objective is maximizing the compression ratio in eq. (2). Then, we let MOBO optimize the rate-distortion tradeoff (i.e., the Pareto frontier surface).

Nonetheless, MOBO is quite complex and requires strong computing capabilities for exploration and exploitation. Particularly, reaching the optimal solution may take days and even weeks, using multiple GPUs. Even on small NN models, to address the high-dimensional search spaces, we apply sparse axis-aligned subspace priors for Bayesian optimization (qNEHVI + SAASBO), with the batch

Noisy Expected Improvement (qNEI) acquisition function, as suggested by Eriksson & Jankowiak (2021); Daulton et al. (2021b;a). Moreover, since the two objectives are not within the same range the cosine similarity objective had to be scaled accordingly to converge to the optimal solution, where a calibration set of 4 images are used during 30 iterations of exploration/exploitation.

Accordingly, we pick a (relatively) small model for comparison (with a size of 112 KB), letting MOBO to find for each layer its optimal bin width and quantize accordingly. We emphasize that the MOBO solution does not rely on the rotation invariant insights. In Figure 4b, the optimization process of MOBO is presented, where each dot depicts experiment results, and its color indicates the iteration in which this result was attained. The compression results of RIQ are presented for comparison. Remarkably, RIQ and MOBO attained almost identical results of $\times 12.6$ and $\times 12.61$, respectively, with a cosine distance of 0.005. This indicates that RIQ reaches the optimal solution.