Proceedings Track

# Do Coresets, Pruning, and Quantization Preserve Neural Network Representations? Exploring Geometry Trajectory Functional Alignment and Representation Similarity [*]

**Editors:** List of editors' names

## Abstract

Neural network compression techniques, such as coreset selection, pruning, and quantization, enable efficient deployment but often induce representational changes that traditional accuracy metrics fail to capture. We propose two complementary and generalizable metrics: Geometry-Trajectory-Functional Alignment (GTFA) and Representation Similarity (REPS). GTFA fuses weight geometry, activation subspace overlap, and confidence-weighted functional similarity, while REPS aggregates effective rank, neuron aliveness, class separation, and eigenvalue decay similarity, providing a multi-faceted evaluation of compression-induced representational degradation. Experiments on CIFAR-10 with ResNet-18 demonstrate that GTFA and REPS correlate strongly with accuracy drops (Pearson $r = 0.806$ and 0.988, respectively), substantially outperforming conventional baselines such as weight similarity ($r = 0.141$) and prediction agreement. Layer-wise visualizations reveal dimensional collapse, neuron death, and class separation degradation, offering interpretable insights into representational integrity under compression. These results position GTFA and REPS as robust, lightweight diagnostic tools for guiding compression-aware model design and adaptive deployment in resource-constrained environments.

**Keywords:** Representation similarity, Neural network compression, Coreset selection, Pruning, Quantization, Weight symmetry and geometry, Accuracy-similarity coupling

## 1. Introduction and Related Work

Deep neural networks (DNNs) achieve state-of-the-art performance across vision, language, and reinforcement learning. Yet, their computational and memory demands limit deployment on resource-constrained platforms such as edge devices, robotics, and IoT Paleyes et al. (2022). This challenge has motivated the development of compression methods that reduce model size and inference cost while maintaining predictive performance, yet their impact on the underlying representational structure remains poorly understood. This gap motivates examining existing compression approaches and their limitations.

**Neural Network and Dataset Compression.** Dataset compression strategies include *coreset selection*, which reduces training data by selecting representative subsets Mirzasoleiman et al. (2020). Whereas, model compression approaches include *pruning*, which eliminates redundant parameters via unstructured $L_1$-norm Han et al. (2015b) or structured filter removal Filters'Importance (2016), and *quantization*, which lowers numerical precision from 32-bit floats to integers or even binary representations Jacob et al. (2018); Courbariaux et al. (2016). Knowledge distillation Hinton et al. (2015) and low-rank factorization Denil et al. (2013) provide complementary model reduction avenues. Despite these

[*] This work was supported by sample footnote.

advances, While these methods improve efficiency, their effect on the *geometry and topology of internal representations* is largely unexplored.

**Neural Network Similarity Metrics.** However, traditional evaluation relies on test accuracy, which provides only a coarse view and misses representational shifts Raghu et al. (2017). Similarity measures such as centered kernel alignment (CKA) Kornblith et al. (2019) compare activation geometry, while subspace overlap methods Vyas et al. (2018) analyze principal components of activation spaces. Functional measures, e.g., KL divergence or prediction agreement, assess output-level consistency Klabunde et al. (2025). Although insightful, these approaches isolate either geometry or function, often failing to explain why compression degrades generalization or robustness. This limitation extends to representation-level analysis.

**Representation Analysis.** Representation-level diagnostics provide richer insights. Intrinsic dimensionality via effective rank Pope et al. (2021) captures complexity, neuron aliveness Morcos et al. (2018b) identifies over-pruning, and class separation Belinkov and Glass (2017) quantifies representational distinctness. Eigenvalue spectrum analysis Ghorbani et al. (2019) characterizes covariance decay. Yet, each measure is fragmented, offering only a partial view of representational collapse under compression, limiting actionable guidance for model design.

**Positioning of GTFA and REPS.** These fragmented perspectives highlight the need for unified metrics. To address these gaps, we introduce two complementary, geometry-aware metrics: *Geometry-Trajectory-Functional Alignment (GTFA):* GTFA combines geometric alignment of activation subspaces across layers with functional fidelity of model outputs. Intuitively, GTFA measures how well compressed networks preserve the representational manifold of hidden states *and* the confidence structure of predictions, offering a joint diagnostic unavailable in prior metrics. Similarly, *Representation Similarity (REPS):* REPS aggregates four complementary representation-level statistics: effective rank, neuron aliveness, class separation, and eigenvalue decay, into a single interpretable score. By unifying multiple structural perspectives, REPS detects subtle collapses in representational geometry that isolated metrics overlook. GTFA extends prior geometric similarity measures Kornblith et al. (2019); Vyas et al. (2018) by unifying representational alignment with functional robustness, while REPS generalizes representation analysis methods Pope et al. (2021); Belinkov and Glass (2017). On CIFAR-10 with ResNet-18, both metrics correlate strongly with accuracy drops ($r = 0.806$ for GTFA and $r = 0.988$ for REPS), substantially outperforming traditional baselines. This demonstrates that geometry- and function-aware diagnostics can guide compression design with high fidelity.

**Our contributions are threefold:**

- We introduce *Geometry-Trajectory-Functional Alignment (GTFA)*, a geometry- and function-aware metric capturing representational perturbations under compression.
- We propose *Representation Similarity (REPS)*, an aggregated diagnostic unifying multiple representation-level statistics into a single interpretable score.
- Through extensive experiments, we show that GTFA and REPS provide robust, efficient, and interpretable tools for guiding compression strategies, uncovering phenomena such as dimensional collapse, neuron death, and degraded class separation.

## 2. Method

### 2.1. Problem Description

Let $(\mathcal{X}, \mathcal{Y})$ denote the input-label spaces and $P$ a distribution on $\mathcal{X} \times \mathcal{Y}$. A reference network $M_0$ with parameters $\theta_0$ implements a measurable map $f_{M_0} : \mathcal{X} \to \Delta^{K-1}$, i.e., a probability distribution over $K$ classes. We evaluate models on a finite test set $S = \{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from $P$. For classification, accuracy is computed as: $\mathrm{Acc}(M; S) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\big[\arg\max f_M(x_i) = y_i\big]$, where the indicator function counts correct predictions. This functional form extends naturally to regression or structured prediction tasks by replacing accuracy with a general metric $\Pi(M; S)$.

**Compression operators.** We consider three compression modalities applied to $M_0$: (i) *coreset selection* (data compression) producing $M_c$, (ii) *pruning* (parameter sparsification) producing $M_p$, and (iii) *quantization* (finite-precision weights/activations) producing $M_q$. Each modality $r \in \mathcal{R}\{c, p, q\}$ is represented by an operator

$$\mathcal{C}_r(\cdot; \lambda) : \mathcal{M} \to \mathcal{M}, \qquad M_{r,\lambda}\mathcal{C}_r(M_0; \lambda), \tag{1}$$

parametrized by a compression budget $\lambda \in \Lambda_r$ (coreset fraction $\alpha$ for $r = c$, sparsity $s$ for $r = p$, bit-width $b$ for $r = q$). where $\lambda$ denotes compression strength: coreset fraction $\alpha \in [0, 1]$, sparsity ratio $s \in [0, 1]$ for pruning, or bit-width $b \in \{1, 2, ..., 8\}$ for quantization, with $\Lambda_r$ representing the respective parameter ranges.unified operator notation emphasizes that heterogeneous compression schemes can be compared on a common geometric footing.

**Geometry-aware model similarity.** Let $\mathcal{L} = \{1, \ldots, L\}$ index layers. For $x \in \mathcal{X}$, $z_\ell(x; M) \in \mathbb{R}^{d_\ell}$ denotes the pre- or post-activation representation at layer $\ell$. With a batch $X_S = [x_1, \ldots, x_n]$, define representation matrices $Z_\ell(M) \in \mathbb{R}^{n \times d_\ell}$ with rows $z_\ell(x_i; M)^\top$ For notational clarity, we use $Z_\ell^{(0)} := Z_\ell(M_0)$ for the original model and $Z_\ell^{(r)} := Z_\ell(M_{r,\lambda})$ for compressed model of type $r$ with parameter $\lambda$. A generic layer similarity $s_\ell : \mathbb{R}^{n \times d_\ell} \times \mathbb{R}^{n \times d_\ell} \to \mathbb{R}$ is assumed invariant to admissible transformations $\mathcal{G}_\ell$:

$$s_\ell\big(Z_\ell^{(0)}, Z_\ell^{(r)}\big) = s_\ell\big(Z_\ell^{(0)} G_1,\, Z_\ell^{(r)} G_2\big) \quad \forall G_1, G_2 \in \mathcal{G}_\ell. \tag{2}$$

where $\mathcal{G}_\ell$ denotes the group of admissible transformations (e.g., orthogonal matrices for rotation invariance). This invariance ensures that similarity measures intrinsic representational geometry (e.g., manifolds, symmetries) rather than coordinate artifacts. For example, an orthogonal basis rotation of hidden activations leaves geometry unchanged but would distort naive similarity measures. Aggregating across layers yields a model-level similarity:

$$\mathrm{MSIM}(M_0, M_{r,\lambda}) = \sum_{\ell \in \mathcal{L}} w_\ell \, s_\ell(Z_\ell(M_0), Z_\ell(M_{r,\lambda})), \quad \sum_\ell w_\ell = 1. \tag{3}$$

where weights $w_\ell$ allow us to emphasize layers that are known to encode semantically rich or symmetry-critical features.

**Accuracy drop.** To quantify compression impact,performance degradation of a compressed model relative to $M_0$ is $\mathrm{ADROP}_{r,\lambda} = \mathrm{Acc}(M_0; S) - \mathrm{Acc}(M_{r,\lambda}; S) \in [0, 1]$. This functional directly measures retained task performance, and its coupling with geometry-aware similarity is central to our analysis.

**Primary objective: similarity-performance coupling.** We investigate whether geometry-preserving compression correlates with retained task performance.

Let $\mathrm{MSIM}_{r,\lambda} := \mathrm{MSIM}(M_0, M_{r,\lambda})$ denote the similarity between original and compressed models. For each modality $r$, we consider the family $\{(\mathrm{MSIM}_{r,\lambda}, \mathrm{ADROP}_{r,\lambda})\}_{\lambda \in \Lambda_r}$ and compute standard correlation measures: Pearson (PLCC), Spearman rank (SRCC), and Kendall's tau (KRCC). For example, SRCC is computed as: $\rho_r = \mathrm{Spearman}\big(\{\mathrm{MSIM}_{r,\lambda}\}_{\lambda \in \Lambda_r}, \{\mathrm{ADROP}_{r,\lambda}\}_{\lambda \in \Lambda_r}\big)$, expecting $\rho_r > 0$ when higher MSIM indicates better representational preservation. We also define a pooled SRCC across all modalities:

$$\rho_{\mathrm{all}} = \mathrm{Spearman}\Big(\{\mathrm{MSIM}_{r,\lambda}\}_{r \in \mathcal{R},\, \lambda \in \Lambda_r}, \{\mathrm{ADROP}_{r,\lambda}\}_{r \in \mathcal{R},\, \lambda \in \Lambda_r}\Big). \tag{4}$$

A strong correlation would provide quantitative evidence that preserving representational geometry under compression directly supports task performance.

**Scope within symmetry & geometry.** Our framework in Equation (3) is agnostic to the choice of $s_\ell$ but requires invariance (Equation (2)) to capture intrinsic geometry. This extends naturally to networks with group symmetries by choosing appropriate invariant similarity functions.

### 2.2. Model Parameter Distribution Analysis

Before defining our specific similarity metrics GTFA and REPS, we first examine how compression affects weight distributions, as these changes directly influence representational geometry. The distribution of learned parameters provides a window into how compression reshapes the inductive biases of the base model $M_0$. **Weight distributions.** Let $\Theta(M)\{\theta_j\}_{j=1}^{|\theta|}$ denote the flattened set of model parameters. The empirical distribution can be expressed as: $p_M(w) = \frac{1}{|\Theta(M)|} \sum_{j=1}^{|\Theta(M)|} \delta(w - \theta_j)$, where $\delta(\cdot)$ is the Dirac delta function. In our implementation, we approximate $p_M(w)$ using 256-bin histograms and reveal how different compression methods distort the parameter space.

We compare the original model $M_0$ against compressed variants: $M_{c,0.5}$ (50% coreset), $M_{p,0.5}$ (50% pruning), and $M_{q,4}$ (4-bit quantization). As shown in Figure 1, different strategies leave distinctive signatures: *Coreset selection* changes variance in the tails due to retraining with reduced data support. *Pruning* creates sharp valley in the distribution near zero, reflecting enforced sparsity. *Quantization* clusters weights into discrete lattice points, with severe information loss when reducing precision from 4-bit to 2-bit. These signatures can be interpreted geometrically: pruning projects weights toward sparse coordinate axes, coreset selection redistributes variance across directions, and quantization discretizes the weight manifold. Each method thus leaves a characteristic "fingerprint" in parameter space, linking geometric perturbations to representational similarity and downstream performance. Understanding these weight-level changes motivates our GTFA metric, which explicitly captures weight geometry alongside activation patterns.

### 2.3. Proposed Model Quality Metrics

To evaluate the quality of compressed neural networks, we propose two complementary metrics: **Geometry-Trajectory-Functional Alignment (GTFA)** and **Representation Similarity (REPS)**. GTFA measures how well compressed models align with the

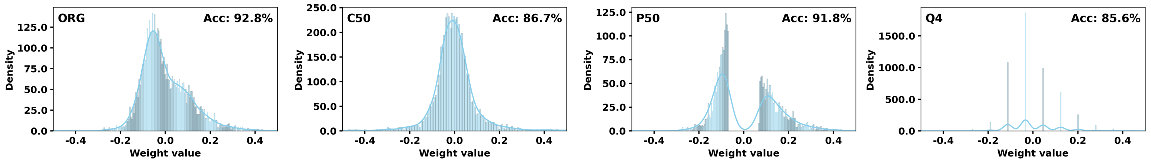


Figure 1: Weight distributions of the original model ORG ($M_0$) and compressed variants: coreset C50($M_{c,0.5}$), pruning P50($M_{p,0.5}$), and quantization Q4($M_{q,4}$). Each compression strategy produces distinct geometric perturbations in parameter space.

original model's weights and activations, while REPS assesses preservation of internal representational structure.

**Geometry-Trajectory-Functional Alignment (GTFA).** GTFA measures alignment between original and compressed models by combining three components: (i) weight similarity $s_w^l$, (ii) activation subspace overlap $s_a^l$, and (iii) confidence-weighted functional similarity $s_f$. The metric builds on CKA Kornblith et al. (2019) and subspace overlap methods Vyas et al. (2018). For a layer $l$, let $\mathbf{W}_0^l$ and $\mathbf{W}_{r,\lambda}^l$ denote flattened weights of the original and compressed models. Normalize as $\hat{\mathbf{W}} = (\mathbf{W} - \mu(\mathbf{W}))/\sigma(\mathbf{W})$ and compute weight similarity:

$$s_w^l = \frac{\hat{\mathbf{W}}_0^l \cdot \hat{\mathbf{W}}_{r,\lambda}^l}{\|\hat{\mathbf{W}}_0^l\|_2 \|\hat{\mathbf{W}}_{r,\lambda}^l\|_2}. \tag{5}$$

For activations, let $\mathbf{A}_0^l, \mathbf{A}_{r,\lambda}^l \in \mathbb{R}^{N \times D_l}$ be activation matrices on test samples. After centering $\tilde{\mathbf{A}} = \mathbf{A} - \bar{\mathbf{A}}$, compute SVDs $\tilde{\mathbf{A}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and define subspace overlap:

$$s_a^l = \frac{1}{k} \sum_{i=1}^{k} \sigma_i(\mathbf{V}_0[:k,:]^\top \mathbf{V}_{r,\lambda}[:k,:]), \qquad k = 20. \tag{6}$$

where $\mathbf{V}_0$ and $\mathbf{V}_{r,\lambda}$ are the top-$k$ right singular vectors from the respective SVDs. This overlap captures whether compressed activations preserve principal subspaces, analogous to comparing neural manifolds across biological recordings Chung and Abbott (2021).

Functional similarity uses confidence-weighted cosine similarity over softmax probabilities, where $\mathbf{P}_{m,i}$ denotes the output probability vector for sample $i$ from model $m$:

$$s_f = \sum_{i=1}^{N} w_i \frac{\mathbf{P}_{0,i} \cdot \mathbf{P}_{r,\lambda,i}}{\|\mathbf{P}_{0,i}\|_2 \|\mathbf{P}_{r,\lambda,i}\|_2}, \quad w_i = \frac{\max(\mathbf{P}_{0,i}, \mathbf{P}_{r,\lambda,i})}{\sum_{k=1}^{N} \max(\mathbf{P}_{0,k}, \mathbf{P}_{r,\lambda,k})}. \tag{7}$$

We first combine weight and activation similarities into a geometric score $s_g^l$, then fuse with functional similarity. The GTFA Fusion is defined as:

$$s_g^l = \beta s_w^l + (1-\beta) s_a^l, \quad \mathrm{GTFA}^l = \gamma s_g^l + (1-\gamma) s_f, \tag{8}$$

with $\beta \in [0,1]$, $\gamma \in [0,1]$. These weights balance structural versus functional alignment. The final GTFA score averages across selected layers (e.g., $\{\text{layer4.1.conv2}, \text{fc}\}$ for ResNet-18).

**Representation Similarity (REPS).** REPS aggregates four complementary neural representation metrics that capture different aspects of representational integrity: effective rank (dimensionality), neuron aliveness (sparsity), class separation (discriminability) and eigenvalue decay similarity (complexity). Each metric is normalized relative to the original model. For activations $\mathbf{A}^l_{r,\lambda}$ at layer $l$:

*Effective Rank* ($r^l_e$): Entropy of normalized singular values $\mathbf{s} = \mathbf{\Sigma}/\|\mathbf{\Sigma}\|_1$:

$$r^l_e = \exp\left(-\sum_{i:s_i>\epsilon} s_i \log s_i\right), \quad \epsilon = 10^{-12}, \quad s^l_e = \min(r^l_e/r^{l,0}_e, 1). \tag{9}$$

*Alive Neurons* ($s^l_{alive}$): Fraction of neurons with variance above threshold:

$$\mathrm{var}_d = \mathrm{Var}(\mathbf{A}^l_{r,\lambda}[:,d]) \quad \bar{v} = \frac{1}{D_l}\sum_d \mathrm{var}_d, \quad s^l_{alive} = 1 - \frac{1}{D_l}\sum_d \mathbb{I}(\mathrm{var}_d < 0.01\bar{v}). \tag{10}$$

*Class Separation* ($r^l_s$): Ratio of inter-class to intra-class distances:

$$\mathbf{c}_c = \frac{1}{N_c}\sum_{i:y_i=c} \mathbf{A}^l_{r,\lambda}[i,:], \quad d_{\text{inter}} = \frac{2}{C(C-1)}\sum_{c<c'} \|\mathbf{c}_c - \mathbf{c}_{c'}\|_2, \tag{11}$$

$$d_{\text{intra}} = \frac{1}{C}\sum_c \frac{1}{N_c}\sum_{i:y_i=c} \|\mathbf{A}^l_{r,\lambda}[i,:] - \mathbf{c}_c\|_2, \quad s^l_s = \min(r^l_s/r^{l,0}_s, 1). \tag{12}$$

*Eigenvalue Decay Similarity* ($s^l_d$): Decay rate $\delta$ from linear fit on log-log plot of top-$m$ eigenvalues $\lambda_i$:

$$\delta = -\text{slope of polyfit}(\log(1:m), \log(\lambda_{1:m})), \quad s^l_d = \frac{1}{1+|\delta-\delta^0|}. \tag{13}$$

REPS combines these metrics using weights: $\text{REPS}^l = w_1 s^l_e + w_2 s^l_{alive} + w_3 s^l_s + w_4 s^l_d$, where $\sum_{i=1}^4 w_i = 1$. We evaluate both weighted and equal-weight variants in our experiments.

## 3. Experimental Setup

This section details the dataset, model architecture, training pipeline, compression protocols, and evaluation metrics used to assess GTFA and REPS on compressed neural networks.

**Dataset.** Experiments are conducted on CIFAR-10 Krizhevsky et al. (2009), consisting of 50,000 training and 10,000 test images of size $32 \times 32$ across 10 classes. CIFAR-10 is selected as it is widely used for compression and generalization studies, enabling reproducible comparisons with prior pruning, quantization, and coreset works Han et al. (2015a); Frankle and Carbin (2018). We apply standard preprocessing procedures using `transforms.Normalize`, and training data is augmented with random horizontal flips and 4-pixel random cropping.

**Model Architecture.** We adopt the ResNet-18 He et al. (2016) architecture. ResNet-18 balances expressivity and computational efficiency, and has been extensively benchmarked in compression studies Mirzasoleiman et al. (2020); Frankle and Carbin (2018).

**Training Implementation Details.** All experiments are implemented in PyTorch on Kaggle with an NVIDIA Tesla P100 GPU. Models are trained from scratch using SGD with momentum 0.9, weight decay $5 \times 10^{-4}$, initial learning rate 0.1, batch size 128, and 50 epochs. Test evaluation uses random seeds fixed at 42. These hyperparameters follow standard CIFAR-10 baselines and ensure reproducibility He et al. (2016); Han et al. (2015a).

**Hyperparameter Settings.** GTFA uses $\beta = 0.45$, $\gamma = 0.45$, and retains the top-$k = 20$ singular vectors. REPS uses $w_1 = 0.15$, $w_2 = 0.05$, $w_3 = 0.35$, $w_4 = 0.45$, $\epsilon = 10^{-12}$, and neuron aliveness threshold 0.01. These values are empirically tuned for stability and consistency across compression regimes, and minor variations do not affect correlation trends.

**Coreset Setup.** Training subsets are selected with fractions [0.8, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01]. Each model is trained from scratch using baseline hyperparameters.

**Pruning Setup.** Global unstructured $L_1$-norm pruning Han et al. (2015a) is applied across all convolutional and fully connected layers, with ratios [0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9]. Post-pruning weights are reparameterized for evaluation without fine-tuning.

**Quantization Setup.** Post-training quantization Jacob et al. (2018) is applied with bit-widths [8, 7, 6, 5, 4, 3, 2, 1]. 8-bit uses PyTorch dynamic quantization on linear layers; lower bit-widths use uniform min-max quantization.

**Evaluation Metrics.** Compressed models are evaluated using: *Classification Accuracy:* Top-1 accuracy on CIFAR-10 test set. *Accuracy Drop:* $\Delta\text{Acc} = \text{Acc}(M_0) - \text{Acc}(M)$, capturing degradation under compression. *Similarity Metrics:* GTFA and REPS computed on representative layers (`layer4.1.conv2`, `fc`) using 200 test samples. *Neural Representation Metrics:* Effective rank, neuron aliveness, class separation, eigenvalue decay similarity. *Functional Metrics:* Prediction agreement, cosine similarity of logits, $L_2$ distance, KL divergence, Wasserstein distance. *Correlation Analysis:* Predictive power of similarity metrics is assessed using PLCC, SRCC, and KRCC between metric values and accuracy drop across all compression settings, capturing both linear and monotonic trends following best practices Kornblith et al. (2019); Neyshabur et al. (2018). Visualizations (CKA trajectories, REPS breakdowns) are generated using `seaborn` and `matplotlib`. These analyses provide quantitative and qualitative insights into how compression alters neural representations.

## 4. Experimental Results

We evaluate twenty-four compressed ResNet-18 models on CIFAR-10 across coreset selection, pruning, and quantization to assess the predictive power of **GTFA** and **REPS**. Our analysis combines quantitative evaluation of test accuracy, correlation metrics (PLCC, SRCC, KRCC), and qualitative visualizations of representational dynamics.

**Accuracy and Compression Effects.** Table 1 summarizes test accuracies under different compression regimes. The uncompressed baseline achieves 0.9277. Coreset compression exhibits smooth accuracy degradation, from 0.9215 at 80% of training data down to 0.1059 with only 1%, reflecting reduced data diversity rather than structural loss. Pruning maintains stability up to 0.5 ratio (0.9177) but drops sharply at 0.7 (0.7992) and 0.9 (0.2336), consistent with removal of critical parameters. Quantization retains high accuracy under moderate bit reduction, e.g., 5-bit (0.9244), but collapses below 4-bit (0.1 at 3-bit and lower). This divergence highlights fundamental differences: pruning selectively removes structured redundancy, whereas low-bit quantization erodes representational precision across all layers.

**Metrics Correlation Comparison.** Table 2 reports PLCC, SRCC, and KRCC correlations with accuracy drop. Weight similarity is weakly correlated (0.1382 PLCC), indicating

Table 1: Test accuracy of ResNet-18 models on CIFAR-10 under coreset, pruning, and quantization. Baseline accuracy: 0.9277.

| **Coreset** | | **Pruning** | | **Quantization** | |
|---|---|---|---|---|---|
| Fraction | Acc. | Ratio | Acc. | Bits | Acc. |
| 0.80 | **0.9215** | 0.05 | **0.9277** | 8 | **0.9281** |
| 0.50 | 0.8673 | 0.10 | 0.9272 | 7 | 0.9276 |
| 0.40 | *0.8649* | 0.20 | *0.9271* | 6 | *0.9254* |
| 0.30 | 0.7887 | 0.30 | 0.9266 | 5 | 0.9244 |
| 0.20 | 0.7286 | 0.40 | 0.9237 | 4 | 0.8558 |
| 0.10 | 0.4668 | 0.50 | 0.9177 | 3 | 0.1000 |
| 0.05 | 0.4717 | 0.70 | 0.7992 | 2 | 0.1000 |
| 0.01 | 0.1059 | 0.90 | 0.2336 | 1 | 0.1000 |

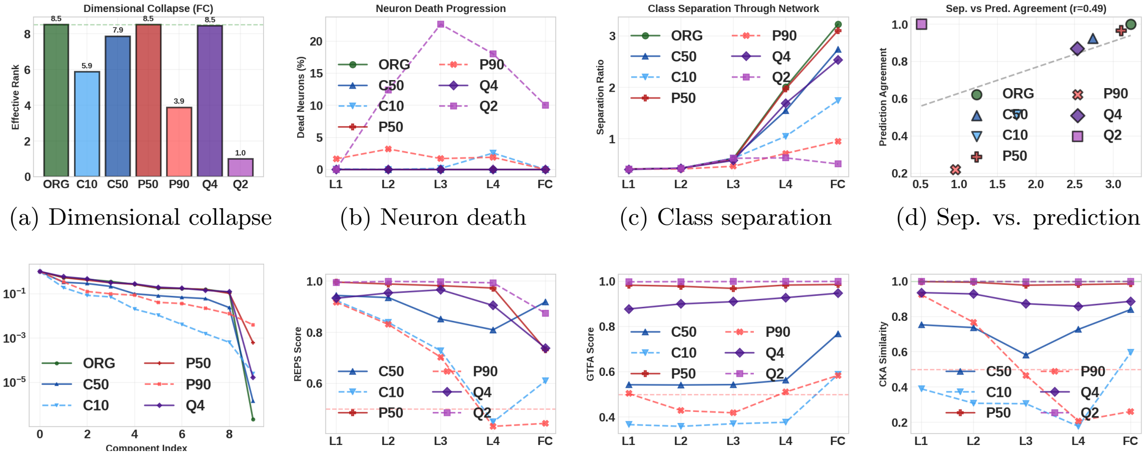
Table 2: Correlations (PLCC, SRCC, KRCC) between similarity metrics and accuracy drop for 24 compressed models.

| **Metric** | **PLCC** | **SRCC** | **KRCC** |
|---|---|---|---|
| Weight Similarity Kornblith et al. (2019) | 0.1382 | 0.4865 | 0.3825 |
| Prediction Agreement Morcos et al. (2018a) | 0.8782 | 0.8028 | *0.7912* |
| Logits Cosine Li et al. (2016) | 0.2344 | -0.0357 | -0.0619 |
| Logits $L_2$ Hinton et al. (2015) | 0.3539 | 0.8407 | 0.8051 |
| KL Divergence Minka et al. (2005) | -0.6736 | -0.8416 | -0.7978 |
| Wasserstein Distance Liu et al. (2025) | -0.4984 | -0.6092 | -0.6011 |
| GTFA | 0.8059 | 0.7868 | 0.7323 |
| Effective Rank | *0.9699* | 0.8269 | 0.7227 |
| Alive Neurons | 0.7162 | 0.6607 | 0.5675 |
| Class Separation | *0.9696* | **0.9408** | **0.8634** |
| Eigenvalue Decay Similarity | 0.8350 | 0.7852 | 0.6168 |
| REPS (weighted) | 0.9735 | 0.8964 | *0.7905* |
| REPS (equal weights) | **0.9878** | *0.8912* | 0.8051 |

limited diagnostic value. Functional metrics show mixed performance: prediction agreement aligns strongly (0.8782 PLCC, 0.8028 SRCC), while logits cosine and $L_2$ provide partial insight, and KL divergence/Wasserstein distances inversely correlate, reflecting instability under compression. GTFA achieves PLCC 0.8059, demonstrating its ability to integrate geometric and activation-based information. REPS delivers superior predictive power: weighted combination yields PLCC 0.9735, while equal weighting improves further to 0.9878, confirming that balanced contributions of effective rank, neuron aliveness, class separation, and eigenvalue decay produce a stable, high-fidelity predictor of accuracy degradation. This demonstrates the theoretical advantage of combining complementary representational indicators rather than relying solely on functional or geometric metrics.

**Qualitative Analysis.** We evaluate representational changes in ResNet-18 model under coreset (C50, C10), pruning (P50, P90), and quantization (Q4, Q2) compression using eight diagnostics (Figure 2). The original model (ORG) achieves a baseline rank of 8.5, 0% neuron death, and 3.2 class separation at FC, while aggressive methods (P90, Q2, C10) show collapsed ranks (e.g., Q2: 1.0), high neuron death (Q2: $\sim$23% in L3), and poor separation (P90: 0.95). Separation correlates strongly with prediction agreement ($r = 0.49$), with P50 (3.1, 0.96) and Q4 (2.5, 0.87) aligning closely with ORG, unlike C10 (1.7, 0.51) and P90 (0.95, 0.22). Eigenvalue decay is steeper for C10 and P90, indicating simplified subspaces, while REPS, GTFA, and CKA trajectories confirm structural and functional divergence in aggressive methods (e.g., P90 CKA: 0.2 at L4). Moderate methods (P50, Q4, C50) preserve integrity, whereas aggressive compression causes collapse, correlating with accuracy drops. See Appendix A for details.

**Discussion.** REPS with equal weighting achieves the highest correlation across all metrics (PLCC 0.9878), surpassing GTFA and functional baselines. This underscores the importance of leveraging complementary neural representation indicators to predict compression-induced accuracy drops. GTFA, while slightly weaker, remains efficient for early diagnostic purposes. Visual and quantitative evidence consistently show that aggressive pruning ($\geq$ 0.70) and ultra-low quantization ($\leq$ 3 bits) lead to severe representational collapse, directly mirroring accuracy loss. These results establish GTFA and REPS as practical, lightweight tools for compression-aware model evaluation, particularly relevant for edge deployments where preserving representational fidelity is critical.



(a) Dimensional collapse (b) Neuron death (c) Class separation (d) Sep. vs. prediction

(e) Eigenvalue decay (f) REPS trajectory (g) GTFA trajectory (h) CKA trajectory

Figure 2: Layer-wise representation analysis across coreset, pruning, and quantization. The eight diagnostics highlight different structural and functional phenomena: (a) dimensional collapse, (b) neuron death, (c) class separation, (d) separation vs. prediction agreement, (e) eigenvalue decay (FC), (f) REPS trajectory, (g) GTFA trajectory, and (h) CKA trajectory.

## 5. Conclusion

We introduced two complementary metrics, **GTFA** and **REPS**, for evaluating compressed neural networks beyond conventional accuracy. Unlike prior approaches that rely solely on weight or prediction similarity, our metrics comprehensively capture structural, functional, and representational fidelity. GTFA integrates weight geometry, activation trajectory, and functional overlap, while REPS aggregates effective rank, neuron aliveness, class separation, and eigenvalue decay, offering a principled measure of representational integrity across layers. On CIFAR-10 with 24 compressed ResNet-18 models, REPS (equal weights) achieves PLCC $= 0.9878$, SRCC $= 0.8912$, and KRCC $= 0.8051$, significantly outperforming GTFA (PLCC $= 0.8059$, SRCC $= 0.7868$, KRCC $= 0.7323$) and conventional similarity metrics. Visual analyses confirm that REPS effectively captures layer-wise representational collapse, while GTFA provides complementary geometric-functional insights. These results establish GTFA and REPS as lightweight, efficient, and generalizable tools for detecting hidden representational degradation and guiding adaptive compression in resource-constrained deployments. Future work will extend these metrics to larger datasets, deeper architectures, and adaptive compression frameworks where layer-wise feedback directly informs compression decisions, enabling automated optimization of accuracy-efficiency trade-offs.

## Acknowledgments

Acknowledgements go here.

## References

Yonatan Belinkov and James Glass. Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, 30, 2017.

SueYeon Chung and Larry F Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70: 137–144, 2021.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando De Freitas. Predicting parameters in deep learning. *Advances in neural information processing systems*, 26, 2013.

Determine Filters'Importance. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

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

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015a.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015b.

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*, pages 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Xinran Liu, Yikun Bai, Yuzhe Lu, Andrea Soltoggio, and Soheil Kolouri. Wasserstein task embedding for measuring task similarities. *Neural Networks*, 181:106796, 2025.

Tom Minka et al. Divergence measures and message passing. 2005.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018a.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018b.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. Challenges in deploying machine learning: a survey of case studies. *ACM computing surveys*, 55(6):1–29, 2022.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European conference on computer vision (ECCV)*, pages 550–564, 2018.

## Appendix A. Representation Analysis

We analyze neural representation changes in ResNet18 model trained on a CIFAR-10 dataset under coreset (C50, C10), pruning (P50, P90), and quantization (Q4, Q2) compression, using eight diagnostic plots (Figure 2).

**Dimensional Collapse** (Figure 2a) shows effective rank at the FC layer. The original model (ORG) has a rank of 8.52. C50 (7.86, 92.3%) and Q4 (8.46, 99.3%) retain high rank, while C10 (5.88, 69.0%) and P90 (3.89, 45.6%) drop significantly. Q2 collapses to 1.00 (11.7%), indicating severe subspace loss. Aggressive compression (P90, Q2) drastically reduces dimensionality, impairing representational capacity, while moderate methods preserve it.

**Neuron Death Progression** (Figure 2b) tracks inactive neurons. ORG, C50, P50, and Q4 show 0.0% death across layers. C10 peaks at 2.6% (L4), and P90 reaches 3.2% (L2). Q2 exhibits high death (22.6% in L3, 10.0% in FC), reflecting quantization artifacts. Q2 and P90 induce significant neuron inactivity, particularly in deeper layers, signaling structural damage.

**Class Separation** (Figure 2c) measures separation ratios. ORG increases from 0.41 (L1) to 3.23 (FC). C50 (2.74, 84.8%) and P50 (3.10, 96.2%) maintain strong separation, while C10 (1.75, 54.1%), P90 (0.95, 29.5%), and Q2 (0.51, 15.9%) collapse in FC. Aggressive methods blur class boundaries, reducing discriminability.

**Separation vs. Prediction Agreement** (Figure 2d) correlates FC separation with prediction agreement ($r = 0.49$). ORG is at (3.23, 1.00), P50 (3.10, 0.97), and Q4 (2.54, 0.87) perform well, while C10 (1.75, 0.52) and P90 (0.95, 0.22) falter. Q2 (0.51, 1.00) is anomalous, suggesting overfitting. Strong correlation validates separation as a performance proxy; P90 and C10 show significant functional loss.

**Eigenvalue Decay** (Figure 2e) shows normalized eigenvalues at FC. ORG decays gradually (1.00, 0.53, . . . , 0.00). C10 (1.00, 0.19, . . . , 0.00) and P90 (1.00, 0.34, . . . , 0.00) decay steeply, indicating simplified subspaces. P50 and Q4 closely follow ORG. Aggressive compression concentrates variance, limiting expressiveness.

**REPS Trajectory** (Figure 2f) tracks REPS scores. P50 (0.99-0.73) and Q4 (0.93-0.74) stay robust, while C10 (0.92-0.61) and P90 (0.92-0.44) drop below 0.50 in FC, reflecting structural collapse. REPS highlights layer-wise sensitivity, with P90 and C10 most affected.

**GTFA Trajectory** (Figure 2g) shows gradient alignment. P50 (0.98-0.99) and Q4 (0.88-0.95) maintain high scores, while C10 (0.37-0.59) and P90 (0.50-0.58) decline. Q2 (1.00-1.00) shows artificially high scores. Low GTFA in C10 and P90 indicates disrupted training dynamics.

**CKA Trajectory** (Figure 2h) measures activation similarity. P50 (0.99-0.99) and Q4 (0.93-0.89) align closely with ORG, while C10 (0.39-0.60) and P90 (0.92-0.26) diverge. Q2 (1.00-1.00) is anomalously high. Low CKA in C10 and P90 confirms functional divergence.

**Overall Analysis.** Aggressive compression (P90, Q2, C10) induces severe representational collapse, evidenced by low rank, high neuron death, poor separation, and functional misalignment, correlating with accuracy drops (e.g., 80.0% for P50). Moderate methods (P50, Q4, C50) preserve representational integrity. REPS and GTFA effectively diagnose compression effects, guiding adaptive strategies for constrained environments.