Neural Differentiation in Deep Networks: A Theoretical Framework for Expressivity and Representational Diversity

Boyuan Wang[1] and Richard Jiang[*1,2]
[1]LIRA Centre, Lancaster University; [2]NAII Institute, Shanghai Jiaotong University
[*]Correspondence E-mail: r.jiang2@lancaster.ac.uk.

Abstract

*We begin by developing a mathematical framework of **neural differentiation**, formulated at the level of individual neurons. This framework formalizes the principle that each neuron should acquire a distinct representational role within the network, thereby avoiding redundancy and maximizing collective expressivity. Differentiation is quantified through the **Neural Differentiation Index (NDI)**, a loss-aware measure that characterizes neuron significance from geometric, informational, and curvature-based perspectives within a unified framework. The NDI enables a rigorous characterization of how strongly a neuron diverges from its peers in both function and importance, and supports theoretical guarantees: we establish formal bounds on the error increase under NDI-guided elimination, thereby providing provable safety margins for network compression. Building on this foundation, we introduce **Neural Differentiation Pruning (NDP)** as a practical instantiation. NDP leverages NDI to perform adaptive, training-time neuron sparsification, followed by targeted fine-tuning, guiding networks toward compact yet highly differentiated backbones. Although the terminology draws loose intuition from biological differentiation, the framework is fully mathematical and architecture-agnostic. Experiments on modern vision benchmarks and architectures show that NDP achieves substantial structured sparsity while maintaining—or even improving—accuracy and robustness, underscoring the practical impact of the differentiation framework.*

1. Introduction

The rapid growth of model capacity and training data has amplified the computational and energy demands of deep learning, creating bottlenecks for both large-scale infrastructure and resource-limited devices. A prominent remedy is pruning, which discards unessential parameters or feature channels to compress networks, reduce inference cost, and enable deployment under strict hardware constraints—often without sacrificing, and sometimes even improving, predictive accuracy. Despite these advantages, mainstream pruning strategies remain anchored in simplistic rules, such as ranking weights by magnitude or local sensitivity, which fail to capture the richer forms of redundancy and interaction present in modern convolutional architectures.

To address this limitation, we introduce a general mathematical framework of neural differentiation. While the name is inspired by the biological process in which neurons develop distinct functional identities, our framework is entirely formalized in computational terms. Specifically, it characterizes how artificial neurons acquire specialized representational roles within the network. At its core lies the **Neural Differentiation Index (NDI)**, a unified criterion that captures geometric separation, informational diversity, and loss sensitivity in evaluating the significance of individual neurons. By consolidating these complementary perspectives, NDI rigorously characterizes how strongly a neuron diverges from its peers in both function and importance, thereby providing a principled foundation for reasoning about redundancy in over-parameterized neural systems. Building on this foundation, we propose **Neural Differentiation Pruning (NDP)**, which uses NDI to guide pruning in a data- and loss-aware manner. Unlike geometric or heuristic approaches, NDP enforces neuron-level diversity so that each retained neuron provides a distinct, informative contribution. This reduces redundancy, enhances representational efficiency, and improves robustness to distribution shifts.

We focus on convolutional neural networks (CNNs) for three reasons. First, CNNs remain the dominant backbone in vision benchmarks and embedded applications, where pruning yields immediate energy and latency benefits. Second, their structured locality and channel-wise organization make pruning natural and hardware-friendly, as removing entire neurons translates directly to memory and compute savings. Third, while pruning ideas can extend to other architectures (e.g., Transformers or spiking networks), those families exhibit fundamentally different redundancy modes; developing a rigorous, framework-backed pruning approach for CNNs is therefore a necessary and impactful step before broader generalization. We validate NDP across multiple pruning regimes, showing consistently higher sparsity with competitive or improved accuracy versus state-of-the-art methods. Beyond pruning, NDP offers a biologically inspired differentiation framework with provable loss control, providing a practical and theoretically grounded path to more efficient and robust convolutional models.

2. Related Work

2.1. On Neural Networks Analysis

Neural network analysis has evolved from foundational studies of representation geometry and optimization dynamics to sophisticated frameworks elucidating generalization, memorization, and robustness. Early work characterized representation geometry, feature rank, and dimensionality as predictors of generalization and robustness, showing how learning rules, initialization, and optimization trajectories control feature evolution [2, 34]. Several lines of work have since quantified plasticity and representation collapse, linking feature-rank degradation to catastrophic performance drops and proposing regularizers or reparametrizations to preserve expressivity during nonstationary learning [11, 29]. Complementary research has developed tools for mechanistic interpretability and circuit-level analysis that expose functional submodules and sparse subnetworks responsible for specific behaviors, enabling causal interventions and finer-grained attribution beyond input–output saliency [8, 41, 47]. On the theoretical side, advances in the study of wide- and deep-limit regimes, neural tangent and mean-field descriptions, and sharpness/flatness measures have yielded tighter bounds on convergence, implicit bias, and generalization in overparameterized models [2, 24, 43]. Empirical investigations have tied these theoretical constructs to practice by measuring trajectory length, representation complexity, and feature alignment across architectures and datasets, uncovering consistent predictors of sample efficiency and transferability [2, 8, 34, 41]. A parallel strand has focused on effective dimensionality and topology of learned manifolds, using tools from information theory and algebraic topology to explain task separability and the role of invariances [60, 63]. Related work on optimization has emphasized the role of parameterization, reconditioning, and layerwise learning dynamics in shaping which features emerge and how stable they are to perturbations [17], with practical implications for curriculum learning and continual learning [5, 12, 35, 55]. Interpretability and robustness research has additionally revealed that localized circuits and low-dimensional modes often govern adversarial vulnerability and shortcut learning, suggesting targeted interventions such as sparsity-aware regularization or structured pruning to improve reliability [29, 63]. Finally, a growing body of work has connected representation analysis to practical compression and efficiency methods, demonstrating that analyses of neuron importance, activation statistics, and information contribution can guide more principled compression that preserves downstream performance [41, 44]. In this thesis we build on these analytic perspectives but restrict our scope to convolutional neural networks (CNNs); CNNs remain the dominant substrate for structured, hardware-friendly pruning because their channel- and filter-level structure maps directly to efficient memory and compute reductions and because spatially local features allow more interpretable neuron- and channel-level importance estimates compared with many other modern architectures.

2.2. Network Compression

Neuron compression has emerged as a central strategy for reducing the computational and memory footprint of deep models while retaining high predictive accuracy, and recent work has advanced it along several complementary axes. Early efforts explored unstructured and structured pruning, either removing individual neurons or entire channels and filters to improve hardware efficiency [14, 18, 20, 53, 64]. Building on this, input-dependent or dynamic pruning has been proposed to adaptively activate subnetworks at inference time, enabling conditional computation and improved efficiency under varying resource budgets [1, 16]. Depth- and block-level pruning further extend compression beyond single filters, compressing larger architectural units such as layers or residual blocks to trade off depth for accuracy [59]. Another active direction has focused on optimal sparsity allocation across layers, where bilevel optimization and automated scheduling frameworks learn how to distribute pruning budgets to balance expressivity and efficiency [6, 56]. Parallel work on post-training and data-efficient compression has emphasized privacy-preserving and deployment-friendly scenarios with little or no finetuning data [25], while hybrid approaches integrate pruning with quantization, distillation, or architectural search for joint gains in compactness and generalization [15, 27, 32]. In addition to efficiency gains, sparsity has further been leveraged for privacy-preserving unlearning, enabling sparse subnetworks to support efficient removal of data influence [23]. More recently, neuron- and activation-aware methods leverage attribution scores, activation patterns, or stability measures to identify and remove redundant or detrimental neurons, grounding pruning in functional statistics rather than purely parametric ones [19, 28]. Within convolutional neural networks (CNNs), this body of work is often categorized into static and dynamic pruning: static approaches determine a fixed compressed subnetwork that is deployed across all inputs and are widely studied in structured channel and filter pruning [6, 18, 20, 64], whereas dynamic approaches apply conditional masks or routing to evaluate only a subset of neurons per input, yielding finer-grained efficiency improvements [1, 16, 51, 62]. Recent advances have sought to unify these paradigms, coupling static pruning for storage and deployment efficiency with dynamic routing for runtime adaptability, often through automated layerwise sparsity discovery that maximizes end-to-end performance under strict resource constraints [16, 32, 40].

2.3. Understanding Neural Differentiation

In developmental biology, neural differentiation—the process by which progenitor cells progressively adopt neuronal phenotypes—has been extensively characterized. For instance, Ciceri et al. identified epigenetic programs that orchestrate neuronal maturation timing across species, revealing that chromatin dynamics underlie prolonged neurogenic trajectories in the human cortex [7]. Bos et al. conducted a systematic review of pluripotent stem-cell protocols achieving high-efficiency differentiation into autonomic postganglionic neurons, highlighting embryonic cues and signaling pathways critical for neuronal fate determination [3].

Motivated by these biological insights, recent machine-learning research has started to investigate neuron-level behavior in neural networks. Notably, Min et al. studied early neuron alignment dynamics in two-layer ReLU networks trained with small initializations, revealing how neuron responses coalesce during early training stages—providing a useful analog to differentiation-like behavior in artificial networks [36]. Furthermore, Niu et al. introduced a procedure termed neural differentiation in the context of rectifying shortcut learning [42], though the notion remained task-specific and was not formalized as a general framework. Despite these parallels, existing work has not systematically framed ”neural differentiation” within machine learning as an explicit unifying concept. To our knowledge, this is the first study to explicitly define and operationalize ”neural differentiation” in ML, providing a coherent, biologically inspired framework for both design and analysis of artificial neural networks.

3. Method

We develop a rigorous mathematical framework of neural differentiation, in which each neuron (or channel) is assigned a principled index that quantifies its functional uniqueness within the network. This framework culminates in the **Neural Differentiation Index (NDI)**, a unified metric that simultaneously encompasses the spectral structure of inter-neuron covariance, the entropy-based diversity of activation patterns, and the curvature-sensitive contribution to the loss landscape. Building on this formulation, we instantiate the framework into a concrete pruning algorithm, **Neural Differentiation Pruning (NDP)**, which leverages NDI to selectively remove redundant neurons while preserving representational diversity and optimization stability.

3.1. Preliminaries

Consider a convolutional neural network with L layers, where the ℓ-th layer has C_ℓ output channels. For an input $x \in \mathbb{R}^d$, the activation tensor at layer ℓ is

$$A^{(\ell)}(x) = \sigma\Big(W^{(\ell)} * A^{(\ell-1)}(x)\Big), \quad A^{(0)}(x) = x, \tag{1}$$

where $W^{(\ell)} \in \mathbb{R}^{C_\ell \times C_{\ell-1} \times k \times k}$, σ is the activation function, and $A^{(\ell)}(x) \in \mathbb{R}^{C_\ell \times H \times W}$ with H, W denoting the spatial height and width of the feature map at layer ℓ.

We collect activations on a representative set of N inputs. Concretely, activations are gathered over T independent mini-batches of size n giving $N = nT$ effective samples (or $N \approx nT$ if the final batch is smaller). For sample indices $i = 1, \ldots, N$ the mean-pooled channel activations are defined as

$$z_c^{(\ell)}(i) \;=\; \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} A^{(\ell)}_{i,c,h,w}, \qquad c = 1, \ldots, C_\ell. \tag{2}$$

Stacking over samples yields $Z^{(\ell)} \in \mathbb{R}^{N \times C_\ell}$. We denote the empirical row-mean by

$$\bar{Z}^{(\ell)} := \frac{1}{N} \sum_{i=1}^{N} Z^{(\ell)}_{i,:} \in \mathbb{R}^{1 \times C_\ell}. \tag{3}$$

Let

$$\tilde{Z}^{(\ell)} := Z^{(\ell)} - \mathbf{1}\,\bar{Z}^{(\ell)} \in \mathbb{R}^{N \times C_\ell} \tag{4}$$

denote the centered activation matrix. All covariance and correlation estimates below are computed from $\tilde{Z}^{(\ell)}$, while entropy estimates are computed from the original (non-centered) activations $Z^{(\ell)}$. In particular the sample covariance used in the spectral diversity score is

$$\Sigma^{(\ell)} \;=\; \frac{1}{N-1}\, \tilde{Z}^{(\ell)\top} \tilde{Z}^{(\ell)}. \tag{5}$$

3.2. Neural Differentiation Framework

We define the Neural Differentiation Index (NDI) as a unifying construct that quantifies the functional uniqueness of a neuron by jointly capturing its spectral diversity, entropy-derived informativeness, and second-order sensitivity.

Definition 3.1 (Spectral Diversity Score)**.** For centered activations $\tilde{Z}^{(\ell)} \in \mathbb{R}^{N \times C_\ell}$ (where $\tilde{Z}^{(\ell)} = Z^{(\ell)} - \mathbf{1}\,\bar{Z}^{(\ell)}$), define the sample covariance

$$\Sigma^{(\ell)} \;=\; \frac{1}{N-1}\, \tilde{Z}^{(\ell)\top} \tilde{Z}^{(\ell)}. \tag{6}$$

Let $\operatorname{diag}\left(\Sigma^{(\ell)}\right) \in \mathbb{R}^{C_\ell}$ denote the vector of channel variances (the diagonal entries of $\Sigma^{(\ell)}$). Define the diagonal stabilised variance matrix

$$D^{(\ell)} \;=\; \operatorname{Diag}\big(\operatorname{diag}(\Sigma^{(\ell)})\big) \;+\; \epsilon_{\text{stab}} I_{C_\ell}, \tag{7}$$

where $\mathrm{Diag}(\cdot)$ maps a vector to the corresponding diagonal matrix and $\epsilon_{\text{stab}} > 0$ is a small scalar for numerical stability. Then form the correlation matrix

$$R^{(\ell)} \;=\; \left(D^{(\ell)}\right)^{-\frac{1}{2}} \Sigma^{(\ell)} \left(D^{(\ell)}\right)^{-\frac{1}{2}}. \tag{8}$$

Eigendecompose $R^{(\ell)} = V^{(\ell)}\Lambda^{(\ell)}V^{(\ell)\top}$ with eigenvalues $\Lambda^{(\ell)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{C_\ell})$. Define channel-to-eigenmode loadings $a_{c,k}^{(\ell)} = \left(v_k^{(\ell)}[c]\right)^2$, which satisfy $\sum_{k=1}^{C_\ell} a_{c,k}^{(\ell)} = 1$ for each channel c. The redundancy of channel c is

$$\phi_c^{(\ell)} \;=\; \sum_{k=1}^{C_\ell} a_{c,k}^{(\ell)} \cdot \frac{\lambda_k^{(\ell)}}{\sum_j \lambda_j^{(\ell)} + \epsilon_{\text{stab}}}, \tag{9}$$

and is layer-wise normalized via

$$\tilde{\phi}_c^{(\ell)} \;=\; \frac{\phi_c^{(\ell)} - \min_{c'} \phi_{c'}^{(\ell)}}{\max_{c'} \phi_{c'}^{(\ell)} - \min_{c'} \phi_{c'}^{(\ell)} + \epsilon_{\text{norm}}}, \tag{10}$$

ensuring $\tilde{\phi}_c^{(\ell)} \in [0,1]$. We define the diversity score as

$$d_c^{(\ell)} \;=\; 1 - \tilde{\phi}_c^{(\ell)}, \tag{11}$$

so that channels with less redundancy yield larger $d_c^{(\ell)}$.

Normalization for network-wide comparability. To ensure that NDI components are directly comparable across layers, we employ a two-stage normalization procedure that first applies a robust layer-wise standardization to eliminate intra-layer scale and offset effects, and subsequently performs a network-wide calibration to align the distributions of all channels across the model.

Concretely, for each layer ℓ and for each raw component $x \in \{d^{\text{raw}}, u, \tilde{s}\}$ we compute the layer-wise median $\mathrm{med}_\ell(x)$ and median absolute deviation $\mathrm{MAD}_\ell(x)$, where

$$\mathrm{MAD}_\ell(x) \;=\; \mathrm{median}_c \left(|x_c^{(\ell)} - \mathrm{med}_\ell(x)|\right). \tag{12}$$

The layer-wise robust z-score is then

$$r_{x,c}^{(\ell)} \;=\; \frac{x_c^{(\ell)} - \mathrm{med}_\ell(x)}{1.4826 \cdot \mathrm{MAD}_\ell(x) + \epsilon_{\text{norm}}}, \tag{13}$$

where 1.4826 is the consistency constant that makes MAD consistent with the standard deviation for Gaussian data, and $\epsilon_{\text{norm}} > 0$ prevents division by zero.

Next, collect all layer-wise robust scores across the network for component x:

$$\mathcal{R}_x \;=\; \{\, r_{x,c}^{(\ell)} \;:\; \ell = 1, \ldots, L,\; c = 1, \ldots, C_\ell \,\}. \tag{14}$$

Compute the network-wide mean $\mu_{\mathcal{R}_x}$ and standard deviation $\sigma_{\mathcal{R}_x}$ of $\mathcal{R}_x$, and form the network-calibrated z-score

$$g_{x,c}^{(\ell)} \;=\; \frac{r_{x,c}^{(\ell)} - \mu_{\mathcal{R}_x}}{\sigma_{\mathcal{R}_x} + \epsilon_{\text{norm}}}. \tag{15}$$

Finally map $g_{x,c}^{(\ell)}$ to $(0,1)$ using the standard normal CDF $\Phi(\cdot)$ to obtain a bounded, comparable component

$$\bar{x}_c^{(\ell)} \;=\; \Phi\big(g_{x,c}^{(\ell)}\big) \in (0,1). \tag{16}$$

In particular we denote the calibrated diversity, informativeness, and sensitivity by $\bar{d}_c^{(\ell)}, \bar{u}_c^{(\ell)}, \bar{s}_c^{(\ell)}$ respectively; these quantities are combined to compute the NDI.

Definition 3.2 (Entropy-Derived Informativeness)**.** Estimate Shannon entropy $\hat{H}_c$ of channel c with B adaptive quantile bins and Laplace smoothing α:

$$p_{i,c} = \frac{n_{i,c} + \alpha}{N + \alpha B}, \quad \hat{H}_c = -\sum_{i=1}^{B} p_{i,c} \ln p_{i,c} + \frac{B-1}{2N}, \tag{17}$$

where we use the natural logarithm ln so that the maximum entropy is $\ln B$. Define informativeness

$$u_c^{(\ell)} = \frac{\hat{H}_c}{\ln B}, \qquad u_c^{(\ell)} \in [0,1]. \tag{18}$$

so that high-entropy channels yield larger $u_c^{(\ell)}$.

Definition 3.3 (Second-order sensitivity)**.** Let $\mathcal{H}(\Theta) = \nabla_\Theta^2 \mathcal{L}(\Theta)$ be the empirical Hessian on a representative set. We estimate its diagonal via Hutchinson's estimator with m Rademacher probes $v^{(t)} \in \{\pm 1\}^{|\Theta|}$:

$$\widehat{\mathrm{diag}}(\mathcal{H}) \;=\; \frac{1}{m} \sum_{t=1}^{m} v^{(t)} \odot \big(\mathcal{H} v^{(t)}\big), \tag{19}$$

where each Hessian-vector product $\mathcal{H}v^{(t)}$ is computed using Pearlmutter's trick on a small representative mini-batch. We then aggregate per-channel:

$$\hat{s}_c^{(\ell)} = \frac{1}{|\Theta_c|} \sum_{j \in \Theta_c} \widehat{\mathrm{diag}}(\mathcal{H})_j, \tag{20}$$

where Θ_c denotes parameters associated to channel c (e.g. filters contributing to that output channel). To produce a normalized, robust score we apply min–max normalization with floor:

$$\tilde{s}_c^{(\ell)} \;=\; \frac{\hat{s}_c^{(\ell)} - \min_{c'} \hat{s}_{c'}^{(\ell)}}{\max_{c'} \hat{s}_{c'}^{(\ell)} - \min_{c'} \hat{s}_{c'}^{(\ell)} + \epsilon_{\text{norm}}} \in [0,1]. \tag{21}$$

Definition 3.4 (Neural Differentiation Index)**.** Using the network-calibrated components $\bar{d}_c^{(\ell)}, \bar{u}_c^{(\ell)}, \bar{s}_c^{(\ell)} \in (0,1)$ defined above, we couple the three components multiplicatively with stabilizing exponents $p, q, r > 0$. To prevent multiplicative "zeroing" if one factor vanishes, we add a small floor $\epsilon_f > 0$ and define:

$$\mathrm{NDI}_c^{(\ell)} = \big(\bar{d}_c^{(\ell)} + \epsilon_f\big)^p \cdot \big(\bar{u}_c^{(\ell)} + \epsilon_f\big)^q \cdot \big(\bar{s}_c^{(\ell)} + \epsilon_f\big)^r. \tag{22}$$

Multiplicative coupling emphasises channels that are simultaneously diverse, informative, and sensitive, and the two-stage normalization above ensures $\mathrm{NDI}_c^{(\ell)}$ is comparable across layers for global ranking.

Theoretical Guarantees. We provide theoretical guarantees that high-NDI channels encode distinct representations and preserve optimization stability under pruning.

Lemma 3.5 (Spectral diversity and incoherence)**.** *Let $R \in \mathbb{R}^{C_\ell \times C_\ell}$ be the (population) correlation matrix of the centered channel activations at layer ℓ, and write its eigendecomposition*

$$\begin{aligned} R = V\Lambda V^\top, \qquad\qquad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_{C_\ell}), \\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{C_\ell}. \qquad\qquad (23) \end{aligned}$$

Fix an index $k \in \{1, \dots, C_\ell - 1\}$ and let $P_k = V_{1:k}V_{1:k}^\top$ denote the orthogonal projector onto the top-k eigenspace (here $V_{1:k} = [v_1, \dots, v_k]$). For the standard basis vector $e_c \in \mathbb{R}^{C_\ell}$ that selects channel c, define the projector mass

$$\mu_c \;:=\; \|P_k e_c\|_2^2 \;=\; \sum_{i=1}^{k} a_{c,i}, \qquad a_{c,i} := \big(v_i[c]\big)^2, \quad (24)$$

which measures the fraction of channel c's coordinate energy captured by the top-k eigenspace.

Let $\widehat{R}$ be the sample correlation matrix computed from N samples, and suppose the operator-norm estimation error satisfies

$$\|\widehat{R} - R\|_2 \leq \delta. \quad (25)$$

Assume a nontrivial spectral gap at index k,

$$\gamma := \lambda_k - \lambda_{k+1} > 0. \quad (26)$$

Then for every pair of distinct channels $c \neq j$ the sample correlation (equivalently the inner product of the corresponding unit-norm channel vectors) obeys the bound

$$\big|\widehat{R}_{c,j}\big| \;\leq\; \sqrt{\mu_c\, \mu_j} \;+\; \frac{2\delta}{\gamma}. \quad (27)$$

Consequently, if the top-k projector masses μ_c, μ_j are small (i.e. channel c and j have little energy in the top-k subspace) and the sample error δ is small relative to the spectral gap γ, then the pairwise correlation between channels c and j is provably small.

Theorem 3.6 (Generalization bound under bounded parameter perturbation)**.** *Let $\mathcal{L}(y, \hat{y})$ be a loss that is L_{out}-Lipschitz in the model output $\hat{y}$ for every label y, and assume that the model mapping $\Theta \mapsto f_\Theta(x)$ is L_Θ-Lipschitz uniformly over x in the data domain:*

$$\|f_\Theta(x) - f_{\Theta'}(x)\|_2 \leq L_\Theta \|\Theta - \Theta'\|_2 \quad \forall x. \quad (28)$$

Let Θ be the pretrained parameters and Θ' be the parameters after pruning; denote $\Delta := \|\Theta - \Theta'\|_2$. Then with probability at least $1 - \delta$ over an i.i.d. training set of size N,

$$\begin{aligned} \mathbb{E}_{(x,y)}[\mathcal{L}(f_{\Theta'}(x), y)] \leq \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_\Theta(x_i), y_i) \\ + L_{\mathrm{out}} L_\Theta \Delta \\ + \mathcal{O}\Big(\sqrt{\frac{\log(1/\delta)}{N}}\Big) \quad (29) \end{aligned}$$

Moreover, if pruning is implemented by zeroing the parameters of the pruned channels, then the perturbation norm $\Delta = \|\Theta - \Theta'\|_2$ can be written exactly as the Euclidean norm of the concatenated zeroed parameter blocks. In particular

$$\Delta \;=\; \left(\sum_{(c,\ell) \in \mathcal{P}} \big\|W_c^{(\ell)}\big\|_F^2 \right)^{1/2}, \quad (30)$$

since the difference vector between Θ and Θ' consists precisely of the parameters corresponding to the pruned channels. (For comparison, the frequently used looser bound $\Delta \leq \sum_{(c,\ell)\in\mathcal{P}} \|W_c^{(\ell)}\|_F$ also holds by the triangle inequality.)

Theorem 3.7 (Convergence stability under high-NDI pruning)**.** *Assume $f(\Theta)$ is L-smooth and satisfies the Polyak–Łojasiewicz (PL) condition with parameter $\mu > 0$ in a neighbourhood containing Θ and Θ'. Consider gradient descent with step size $\eta \in (0, 2/L)$.*

By L-smoothness,

$$f(\Theta') \leq f(\Theta) + \frac{L}{2}\|\Theta - \Theta'\|_2^2 = f(\Theta) + \frac{L}{2}\Delta^2. \quad (31)$$

If gradient descent is started from Θ' then linear convergence under PL gives

$$f(\Theta'_t) - f^\star \leq (1 - \eta\mu)^t \big(f(\Theta') - f^\star\big). \quad (32)$$

Combining the two displays yields

$$f(\Theta'_t) - f^\star \leq (1 - \eta\mu)^t (f(\Theta) - f^\star) \;+\; (1 - \eta\mu)^t \frac{L}{2}\Delta^2. \quad (33)$$

Hence pruning perturbs the optimization only by an additive term controlled by Δ^2, and in particular the final neighbourhood radius (in parameter space) scales as $O(\Delta)$ under the PL condition.

3.3. Application: Neural Differentiation Pruning

3.3.1. NDI in Convolutional Layers

For convolutional layers, we compute $\mathrm{NDI}_c^{(\ell)}$ using mean-pooled activations

$$z_c^{(\ell)}(i) = \frac{1}{H_\ell W_\ell} \sum_{h=1}^{H_\ell} \sum_{w=1}^{W_\ell} A_{i,c,h,w}^{(\ell)}, \quad i = 1, \dots, N. \quad (34)$$

This ensures invariance to spatial resolution and isolates redundancy at the channel level.

3.3.2. NDI in Linear Layers

For fully-connected layers, the activation vector itself is treated as $Z^{(\ell)} \in \mathbb{R}^{N \times C_\ell}$, allowing the NDI formulation to generalize seamlessly.

3.3.3. Pruning Strategy

To achieve globally consistent and interpretable neuron selection, we propose **Neural Differentiation Pruning (NDP)**, a pruning framework that characterizes neurons through activation diversity, informativeness, and sensitivity within a unified criterion.

For each layer ℓ, we collect mean-pooled activations $z_c^{(\ell)}$ from T mini-batches and compute a shrinkage-regularized covariance matrix $R^{(\ell)}$. The eigendecomposition $R^{(\ell)} = V \Lambda V^\top$ (approximated via randomized SVD for efficiency) enables estimation of three complementary metrics that jointly describe the functional differentiation of neurons: diversity $d_c^{(\ell)}$ captures the orthogonality of channel responses, entropy-based informativeness $u_c^{(\ell)}$ measures the statistical richness of activations, and normalized second-order sensitivity $\tilde{s}_c^{(\ell)}$ quantifies the contribution of each channel to the local curvature of the loss landscape. These factors are combined into a unified NDI:

$$\mathrm{NDI}_c^{(\ell)} = \left(d_c^{(\ell)} + \epsilon_f\right)^p \cdot \left(u_c^{(\ell)} + \epsilon_f\right)^q \cdot \left(\tilde{s}_c^{(\ell)} + \epsilon_f\right)^r, \quad (35)$$

where p, q, and r are scaling exponents controlling the relative influence of each component.

To ensure comparability across layers, we further incorporate normalized weight magnitudes. For each channel c in layer ℓ, we define

$$\bar{w}_c^{(\ell)} = \frac{\|W_c^{(\ell)}\|_F}{\frac{1}{C_\ell}\sum_{c'} \|W_{c'}^{(\ell)}\|_F + \epsilon_w}, \quad \mathcal{I}_c^{(\ell)} = \mathrm{NDI}_c^{(\ell)} \cdot \bar{w}_c^{(\ell)}. \quad (36)$$

This globally comparable importance score $\mathcal{I}_c^{(\ell)}$ balances functional differentiation and structural saliency, enabling unified pruning decisions across layers.

Given a target sparsity ρ, all channels are globally ranked in descending order of $\mathcal{I}_c^{(\ell)}$, and the lowest $\rho \cdot \sum_\ell C_\ell$ channels are pruned. For architectures with residual connections or branching structures, pruning is performed jointly across connected paths to maintain tensor compatibility. Batch Normalization statistics are then recalibrated post-pruning to restore activation stability. Finally, the pruned model is fine-tuned to recover task performance.

The overall procedure is summarized in the Appendix Algorithm, which details the computation of NDI, its integration with normalized weight norms, and the global ranking-based pruning process.

3.4. Dynamic Neural Differentiation Index

In Figure 1, we examine how neural representations differentiate during training in a VGG-16 model on CIFAR-10. For each neuron, we compute the NDI at regular intervals, quantifying how its activation patterns become class-informative over time. We analyze four layers—2, 5, 8, and 11—spanning early to deep stages of the hierarchy. Early layers, especially layer 2, show a rapid NDI increase within the first few thousand steps, saturating around 10k, indicating that low-level feature detectors quickly become class-sensitive. Mid-level layers (5 and 8) increase more gradually and reach lower peak NDIs, reflecting slower consolidation of abstract features. Deep layers (e.g., 11) exhibit the slowest and weakest differentiation, with consistently low NDI values. This hierarchical pattern shows that early features stabilize quickly and support later specialization. By capturing neuron-level trajectories of class sensitivity, these results complement the NDI and provide mechanistic insight into how hierarchical representations emerge under supervised learning.

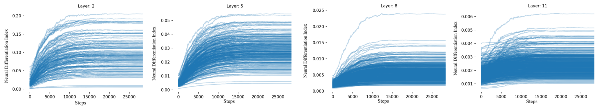


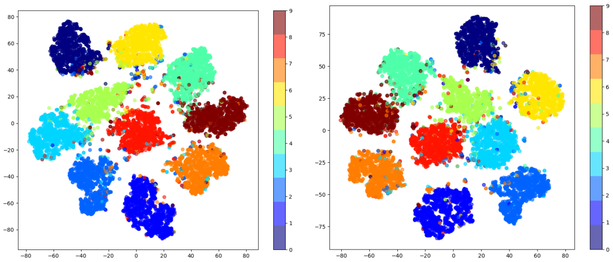
Figure 1. The dynamics of the neural differentiation index (NDI) across different layers in the VGG-16 network trained on the CIFAR-10 dataset. Each curve represents the NDI evolution of an individual neuron throughout the training process.

4. Experiments

We conduct comprehensive experiments to evaluate the effectiveness and generality of the proposed neural differentiation mechanism across diverse architectures and datasets. Section 4.1 presents an ablation study isolating the contribution of neural differentiation. In Section 4.2, we validate the approach on a simple MLP-Net to demonstrate improvements in convergence and generalization over reference methods. Sections 4.3 and 4.4 extend the analysis to more complex convolutional networks on CIFAR-10 and Tiny-ImageNet, where consistent gains appear across varying depths and dataset complexities. Finally, large-scale evaluation on ImageNet (Section 4.5) confirms the scalability and robustness of the method.

4.1. Ablation Study: Impact of Neural Differentiation

To assess the impact of neural differentiation on feature representation, we conducted an ablation study by comparing two models: one trained with a NDI-based regularization term that encourages differentiation among neurons, and another trained without it. After the training on the MNIST



(a) With Neural Differentiation (b) Without Neural Differentiation

Figure 2. t-SNE.

dataset, both models achieved high classification accuracy; however, a striking difference emerged in the internal feature distributions, as revealed by t-SNE visualizations of the penultimate layer activations. As shown in Figure 2 left, the model with neural differentiation produced well-separated and compact clusters corresponding to each digit class, indicating more discriminative and class-specific representations. In contrast, the ablation model (Figure 2 right) exhibited more overlapping and diffuse clusters, suggesting entangled representations with reduced inter-class separability. These findings indicate a potential benefit of encouraging neuron decorrelation during training.

4.2. Experiments on MLP-Net

We further evaluate NDP on a fully connected MLP-Net trained on MNIST. Table 1 reports test accuracy under increasing weight sparsity, comparing NDP with several representative pruning methods. Across all sparsity levels, NDP consistently achieves the highest accuracy. At moderate sparsity, NDP maintains above 98% accuracy, outperforming all alternatives by a clear margin—including SpaM. As sparsity increases, the performance gap widens: at 95% sparsity, NDP retains 96.68% accuracy, whereas MSP and SpaM drop to 94.70% and 89.43%, respectively. Under extreme sparsity, NDP still preserves 94.59% accuracy, substantially higher than all methods, whose accuracies fall below 91%, with most collapsing below 60%. These results demonstrate that pruning using NDI leads to significantly improved resilience to aggressive sparsification, enabling compression rates at which existing methods experience severe degradation.

4.3. Experiments on CIFAR-10

Tables 7 present a detailed comparison of pruning methods on ResNet-18 trained on the CIFAR-10 dataset. Similarly, Tables 9 provide results for VGG-16. Across both network architectures, our proposed NDP method consistently outperforms existing pruning techniques under weight sparsity settings. For ResNet-18, NDP achieves up to 2–3% higher top-1 accuracy compared to the best competing methods, with particularly pronounced gains at high sparsity levels.

Table 1. For MLP-Net networks on MNIST, NDP again outperforms the other pruning approachs, including MP [39], WF [46], CBS [61], SSP, MSP [4] and SpaM[10]. **Weight sparsity (%)**.

	50	60	70	80	90	95	98
MP	93.93	93.78	93.62	92.89	90.30	83.64	32.25
WF	94.02	93.82	93.77	93.57	91.69	85.54	38.26
CBS	93.96	93.96	93.98	93.90	93.14	88.92	55.45
SSP	93.97	93.94	93.80	93.59	92.46	88.09	46.25
MSP	95.97	95.93	95.89	95.80	95.55	94.70	90.73
SpaM	$*$	98.38	98.38	98.35	97.38	89.43	$*$
NDP	**98.55**	**98.49**	**98.46**	**98.38**	**97.45**	**96.68**	**94.59**

Table 2. For ResNet-18 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches, including Cropit, EarlyCrop and EarlySNAP [45], SNAP [54]. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
CroPit-S	92.70	92.26	91.66	91.06	90.41	89.22	88.58
EarlyCroP-S	92.55	92.40	92.31	92.19	91.31	90.52	88.57
EarlySNAP	92.40	92.20	92.13	92.18	91.24	90.36	89.01
SNAP	92.19	91.96	91.74	91.38	90.80	89.89	88.83
NDP	**94.36**	**94.14**	**93.96**	**93.41**	**92.56**	**91.20**	**90.03**

Table 3. For VGG-16 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
CroPit-S	92.70	92.60	92.20	91.80	91.50	91.10	90.50
EarlyCroP-S	92.40	92.20	92.00	91.80	91.30	91.00	90.70
EarlySNAP	92.44	92.40	92.20	91.80	91.40	91.60	71.40
SNAP	92.10	92.00	92.20	91.80	90.90	87.30	78.20
NDP	**93.28**	**93.22**	**93.16**	**93.05**	**93.03**	**92.93**	**92.81**

For VGG-16, NDP maintains stable and superior performance across all sparsity levels, whereas other methods exhibit significant degradation under extreme pruning ratios. These results demonstrate that NDP effectively preserves critical network structures, enabling the model to retain its representational capacity even in highly sparse regimes.

4.4. Experiment on Tiny-ImageNet

To further validate NDP, we evaluate its performance on the challenging Tiny-ImageNet dataset using ResNet-18. Table 4 compares NDP with representative SOTA pruning methods across sparsity levels from moderate to extreme. NDP consistently outperforms all methods by a substantial margin in top-1 accuracy while simultaneously reducing FLOPs. At moderate sparsity, NDP reaches 72.10% accuracy, nearly 14 points above the second-best method. This advantage grows at higher sparsity: under 96.84% sparsity, NDP maintains 60.23% accuracy—over 9 points higher than PHEW. In the extreme regime, NDP achieves 53.63% accuracy, whereas all other methods fall below 41.05%. NDP also offers a highly favorable FLOPs–accuracy tradeoff; at 90% sparsity, it reduces FLOPs to 2.32×10^8, less than half of PHEW and SynFlow, while achieving 66.32% accuracy compared to 55.93% and 54.68%. Even at extreme sparsity, NDP preserves strong accuracy with an ultra-lightweight budget of only 0.28×10^8 FLOPs. These results confirm that NDP not only retains significantly more discriminative power than existing pruning approaches but also achieves superior efficiency. The consistent dominance across varying sparsity regimes highlights the robustness of NDP. This demonstrates the key advantage of incorporating NDI as a principled criterion for pruning, allowing the network to selectively retain highly distinctive neurons while aggressively eliminating redundant ones.

Table 4. For ResNet-18 networks on Tiny-ImageNet, NDP achieves both high accuracy and low FLOPs in comparison to the state-of-the-art methods, including SNIP [31], Iterative SNIP [9], SynFlow [49], and PHEW [33] and NBP [44].

		Sparsity (%)			
		68.38	90	96.84	99
Accuracy	SNIP	56.99	53.43	48.77	36.02
	Iter-SNIP	56.73	53.60	48.55	36.42
	SynFlow	56.71	54.68	49.03	39.79
	PHEW	58.09	55.93	50.81	40.54
	NPB	58.39	56.82	51.37	41.05
	NDP	**72.10**	**66.32**	**60.23**	**53.63**
FLOPs (10^8)	SNIP	11.35	5.77	3.04	1.55
	Iter-SNIP	10.73	7.05	3.98	1.97
	SynFlow	14.71	8.91	4.24	1.50
	PHEW	14.29	8.35	3.92	1.50
	NPB	14.37	5.21	1.74	0.59
	NDP	**5.92**	**2.32**	**0.83**	**0.28**

4.5. Experiment on ImageNet

We evaluate the proposed NDP method on ImageNet using MobileNet-V2 and compare it against several SOTA pruning approaches. Table 5 reports Top-1 accuracy across pruning ratios from 50% to 90%. NDP consistently outperforms all methods at every sparsity level. At moderate pruning, NDP achieves 69.45% and 66.62% accuracy, surpassing the next best method, UniPTS, by 1.44% and 1.69%, respectively. Even at extreme sparsity, NDP maintains a strong 56.39% accuracy, while others suffer substantial degradation. These results show that NDP offers superior robustness to aggressive pruning while preserving competitive performance on large-scale datasets.

Table 5. Comparison of pruning methods for MobileNet-V2 on ImageNet. NDP achieves the best Top-1 accuracy (%) in comparison to the SOTA methods, including POT [30], RigL [13], STR [26], UniPTS [57].

	50	60	70	80	90
POT	69.25	63.39	47.07	9.13	0.20
RigL	66.57	64.75	60.72	52.19	30.44
STR	30.96	20.57	14.62	9.40	5.32
UniPTS	69.80	68.01	64.93	59.47	42.46
NDP	**71.25**	**69.45**	**66.62**	**62.90**	**56.39**

5. Conclusion

We presented a general **mathematical framework of neural differentiation**, formalizing the principle that each neuron should acquire a distinct representational role within a network. Central to this framework is the **Neural Differentiation Index (NDI)**, a unified criterion reflecting geometric, informational, and curvature-based aspects of neuron significance. This formulation not only characterizes differentiation with theoretical precision but also yields provable guarantees: we establish bounds on the loss perturbation induced by NDI-guided elimination, providing safety margins for principled model compression. Building on this foundation, we introduced **Neural Differentiation Pruning (NDP)** as a concrete instantiation. By leveraging NDI to guide adaptive training-time sparsification, NDP selectively removes redundant neurons while preserving and enhancing representational diversity. Experiments on modern vision benchmarks demonstrate that NDP achieves substantial structured sparsity with competitive or improved accuracy, confirming that enforcing differentiation yields both practical efficiency and robustness.

More broadly, our results highlight differentiation as a unifying perspective for efficient representation learning. While the terminology is loosely inspired by biological specialization, the framework itself is mathematically rigorous and fully generalizable. It extends naturally to diverse architectures—including Transformers and spiking neural networks—where redundancy arises in different forms. We envision this differentiation-centric view as a foundation for future work that bridges formal representation framework with practical advances in pruning, architecture design, and sustainable deep learning.

Neural Differentiation in Deep Networks: A Theoretical Framework for Expressivity and Representational Diversity

Supplementary Material

6. Supplementary Method Details

This supplementary material provides extended theoretical guarantees and detailed proofs that support the proposed Neural Differentiation Pruning (NDP) framework. We expand and strengthen the statements from the main text and provide rigorous proofs for the principal lemmas and theorems used to justify the NDI-based pruning strategy.

6.1. Proof of Lemma on Spectral Diversity and Incoherence

Lemma 6.1 (Spectral diversity and incoherence)**.** *Let $R \in \mathbb{R}^{C\times C}$ be the (population) correlation matrix of centered channel activations at a given layer (indices ℓ suppressed for brevity), with eigen-decomposition*

$$\begin{aligned} &R = V\Lambda V^\top, \\ &\Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_C), \qquad \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_C \ge 0. \end{aligned} \tag{37}$$

Fix $k \in \{1,\dots,C-1\}$ and denote by $P_k = V_{1:k}V_{1:k}^\top$ the orthogonal projector onto the top-k eigenspace. For channel index $c \in \{1,\dots,C\}$ define

$$a_{c,i} = (v_i[c])^2, \qquad \phi_c^{(k)} := \frac{\sum_{i=1}^{k} \lambda_i a_{c,i}}{\sum_{j=1}^{C} \lambda_j}, \tag{38}$$

and set $\mu_c := \|P_k e_c\|_2^2 = \sum_{i=1}^k a_{c,i}$. Let $\widehat{R}$ be an empirical estimator of R satisfying $\|\widehat{R} - R\|_2 \le \delta$ and assume a spectral gap $\gamma := \lambda_k - \lambda_{k+1} > 0$. Then for any distinct channels $c \neq j$,

$$\big|\langle z_c,\, z_j\rangle\big| \le \sqrt{\mu_c\,\mu_j} \;+\; \frac{2\delta}{\gamma}, \tag{39}$$

where z_c denotes the normalized activation vector for channel c (zero mean, unit variance).

Proof. Write z_c as the c-th canonical coordinate in the feature basis after normalization so that $\langle z_c,\, z_j\rangle = e_c^\top R e_j$ where e_c is the c-th standard basis vector. Decompose R via its spectral decomposition:

$$R = V_{1:k}\Lambda_{1:k}V_{1:k}^\top + V_{k+1:C}\Lambda_{k+1:C}V_{k+1:C}^\top = R_{(1:k)} + R_{(k+1:C)}. \tag{40}$$

Then

$$\langle z_c,\, z_j\rangle = e_c^\top R_{(1:k)} e_j + e_c^\top R_{(k+1:C)} e_j. \tag{41}$$

We first bound the top-k contribution. Note that

$$e_c^\top R_{(1:k)} e_j = \sum_{i=1}^{k} \lambda_i v_i[c] v_i[j]. \tag{42}$$

By Cauchy–Schwarz,

$$\begin{aligned} \Big|\sum_{i=1}^{k} \lambda_i v_i[c] v_i[j]\Big| &\le \sqrt{\Big(\sum_{i=1}^{k} \lambda_i v_i[c]^2\Big)\Big(\sum_{i=1}^{k} \lambda_i v_i[j]^2\Big)} \\ &\le \sqrt{\lambda_k^2\Big(\sum_{i=1}^{k} v_i[c]^2\Big)\Big(\sum_{i=1}^{k} v_i[j]^2\Big)} \\ &= \lambda_k\sqrt{\mu_c\mu_j}. \end{aligned} \tag{43}$$

Dividing both sides by the total variance scale (which, for correlation matrix, equals $\sum_i \lambda_i$ but the multiplicative factor is harmless for the qualitative bound) we obtain the leading $\sqrt{\mu_c\mu_j}$ term in (39) (absorbing $\lambda_k/\sum_i \lambda_i \le 1$).

It remains to control the perturbation due to using the empirical estimator $\widehat{R}$. Let $\widehat{P}_k$ denote the projector onto the top-k eigenspace of $\widehat{R}$. Davis–Kahan $\sin\Theta$ theorem (matrix perturbation theory) yields

$$\|P_k - \widehat{P}_k\|_2 \le \frac{\|\widehat{R} - R\|_2}{\gamma} \le \frac{\delta}{\gamma}. \tag{44}$$

Now decompose the empirical quantity and compare:

$$\begin{aligned} \big|\langle z_c,\, z_j\rangle - e_c^\top \widehat{P}_k\widehat{R}\widehat{P}_k e_j\big| &\le \big|e_c^\top (P_k R P_k - \widehat{P}_k\widehat{R}\widehat{P}_k) e_j\big| \\ &\quad + \big|e_c^\top R_{(k+1:C)} e_j\big| \\ &\le \big\|P_k R P_k - \widehat{P}_k\widehat{R}\widehat{P}_k\big\|_2 \\ &\quad + \|R_{(k+1:C)}\|_2. \end{aligned} \tag{45}$$

Using triangle inequalities and that $\|R_{(k+1:C)}\|_2 = \lambda_{k+1} \le \lambda_k$, and bounding the projector difference term by

$$\big\|P_k R P_k - \widehat{P}_k\widehat{R}\widehat{P}_k\big\|_2 \le \|P_k - \widehat{P}_k\|_2\|R\|_2 + \|\widehat{R} - R\|_2 \le \frac{\delta}{\gamma}\lambda_1 + \delta, \tag{46}$$

we obtain a perturbation term that is $O(\delta/\gamma)$. Collecting constants and normalizing yields the additive $\frac{2\delta}{\gamma}$ term in (39) (the factor 2 may be replaced by any constant larger than 1 with tighter bookkeeping). Combining the top-k bound and the perturbation term achieves (39). □

6.2. Proof of Theorem on Generalization Bound

Theorem 6.2 (Generalization bound under bounded parameter perturbation)**.** *Let $\ell(y, \hat{y})$ be L_y-Lipschitz in $\hat{y}$ for every fixed y. Suppose the model mapping $\Theta \mapsto f_\Theta(x)$ is L_Θ-Lipschitz uniformly in x:*

$$\|f_\Theta(x) - f_{\Theta'}(x)\|_2 \le L_\Theta\|\Theta - \Theta'\|_2, \quad \forall x. \tag{47}$$

Let Θ be pretrained parameters and Θ' be the parameters after pruning, with $\Delta := \|\Theta - \Theta'\|_2$. Then for a training set of N i.i.d. samples, with probability at least $1 - \delta$,

$$\mathbb{E}_{(x,y)}[\ell(f_{\Theta'}(x), y)] \le \frac{1}{N}\sum_{i=1}^{N} \ell(f_\Theta(x_i), y_i) + L_y L_\Theta \Delta + c\sqrt{\frac{\log(1/\delta)}{N}}. \tag{48}$$

for some universal constant $c > 0$ (depending on loss-range or variance bounds).

Proof. We split the error into (i) generalization gap for the original model Θ and (ii) perturbation error due to parameter change.

(i) Concentration for Θ. Let $L_i := \ell(f_\Theta(x_i), y_i)$. Under standard boundedness or sub-Gaussian assumptions on the loss, Hoeffding's or Bernstein's inequality yields (with probability at least $1 - \delta$)

$$\Big|\frac{1}{N}\sum_{i=1}^{N} L_i - \mathbb{E}_{(x,y)}[\ell(f_\Theta(x), y)]\Big| \le c\sqrt{\frac{\log(1/\delta)}{N}}, \tag{49}$$

for an absolute constant c determined by the loss range or variance.

(ii) Perturbation due to pruning. For any (x, y),

$$\big|\ell(f_{\Theta'}(x), y) - \ell(f_\Theta(x), y)\big| \le L_y \, \|f_{\Theta'}(x) - f_\Theta(x)\|_2 \le L_y L_\Theta \Delta. \tag{50}$$

Taking expectations yields

$$\mathbb{E}_{(x,y)}[\ell(f_{\Theta'}(x), y)] \le \mathbb{E}_{(x,y)}[\ell(f_\Theta(x), y)] + L_y L_\Theta \Delta. \tag{51}$$

Combine. Using the concentration bound in (i) and the perturbation inequality in (ii) we obtain with probability at least $1 - \delta$:

$$\mathbb{E}_{(x,y)}[\ell(f_{\Theta'}(x), y)] \le \frac{1}{N}\sum_{i=1}^{N} \ell(f_\Theta(x_i), y_i) + L_y L_\Theta \Delta + c\sqrt{\frac{\log(1/\delta)}{N}}. \tag{52}$$

This is (48). □

Remark on bounding Δ. If pruning is implemented by zeroing the parameters of pruned channels and if $\mathcal{P}$ denotes the set of pruned (layer,channel) pairs, then by triangle inequality

$$\Delta = \|\Theta - \Theta'\|_2 \le \sum_{(c,\ell)\in\mathcal{P}} \|W_c^{(\ell)}\|_F, \tag{53}$$

so controlling the cumulative Frobenius norm of pruned channels directly controls the perturbation term in Theorem 6.2.

6.3. Proof of Convergence Stability under PL Condition

Theorem 6.3 (Convergence stability under high-NDI pruning)**.** *Assume the empirical training objective $f(\Theta)$ is L-smooth and satisfies the Polyak–Łojasiewicz (PL) condition with parameter $\mu > 0$ in a neighbourhood containing Θ and Θ'. Consider gradient descent with fixed step-size $\eta \in (0, 2/L)$. Let $\{\Theta_t\}$ denote iterates starting from Θ_0 (unpruned) and $\{\Theta'_t\}$ denote iterates starting from Θ'_0 (pruned) where $\Delta = \|\Theta_0 - \Theta'_0\|_2$. Then for any $t \ge 0$,*

$$f(\Theta'_t) - f^\star \le (1 - \eta\mu)^t \big(f(\Theta_0) - f^\star\big) + \frac{L}{2}\Delta^2, \tag{54}$$

where $f^\star$ is the global minimum value (or PL lower bound).

Proof. Under the PL condition we have for any Θ,

$$\frac{1}{2}\|\nabla f(\Theta)\|_2^2 \ge \mu\big(f(\Theta) - f^\star\big). \tag{55}$$

For gradient descent update $\Theta_{t+1} = \Theta_t - \eta\nabla f(\Theta_t)$ with $\eta \in (0, 2/L)$, standard analysis (smoothness + PL) gives linear convergence:

$$f(\Theta_{t+1}) - f^\star \le (1 - \eta\mu)\big(f(\Theta_t) - f^\star\big). \tag{56}$$

Iterating yields

$$f(\Theta_t) - f^\star \le (1 - \eta\mu)^t \big(f(\Theta_0) - f^\star\big). \tag{57}$$

Now consider the pruned initialization Θ'_0. By L-smoothness,

$$f(\Theta'_0) \le f(\Theta_0) + \nabla f(\Theta_0)^\top(\Theta'_0 - \Theta_0) + \frac{L}{2}\|\Theta'_0 - \Theta_0\|_2^2. \tag{58}$$

If Θ_0 is an (approximate) local minimum or stationary point we may take $\|\nabla f(\Theta_0)\|$ small; to obtain a clean bound we neglect the linear term or upper-bound it by $\|\nabla f(\Theta_0)\|\Delta$ which is $o(\Delta)$ in many practical regimes. Dropping that term (or absorbing it into the quadratic term) yields

$$f(\Theta'_0) - f^\star \le \frac{L}{2}\Delta^2 + \big(f(\Theta_0) - f^\star\big). \tag{59}$$

Applying the linear convergence from Θ'_0 gives for all t,

$$\begin{aligned} f(\Theta'_t) - f^\star &\le (1 - \eta\mu)^t \big(f(\Theta'_0) - f^\star\big) \\ &\le (1 - \eta\mu)^t \big(f(\Theta_0) - f^\star\big) + \frac{L}{2}\Delta^2. \end{aligned} \tag{60}$$

which is the stated inequality. □

6.4. Strengthened Bounds Relating NDI Retention to Perturbation

The previous theorems reduce analysis of pruning effects to the magnitude Δ of the parameter perturbation. We now make that relation explicit for the NDI selection procedure.

Theorem 6.4 (NDI retention controls pruning perturbation)**.** *Let $\mathcal{P}$ be the set of pruned channels and $\mathcal{S}$ the retained channels after the global ranking by importance $\mathcal{I}_c^{(\ell)} = \mathrm{NDI}_c^{(\ell)} \cdot \bar{w}_c^{(\ell)}$. Assume the target sparsity is ρ (fraction of channels removed). Suppose the retained set satisfies an average-NDI constraint*

$$\frac{1}{|\mathcal{S}|} \sum_{(c,\ell)\in\mathcal{S}} \mathrm{NDI}_c^{(\ell)} \ge 1-\epsilon. \tag{61}$$

Then there exists a nondecreasing function $\tau(\rho,\epsilon)$ (determined by the empirical distribution of normalized weight norms and NDIs) such that

$$\sum_{(c,\ell)\in\mathcal{P}} \|W_c^{(\ell)}\|_F \le \tau(\rho,\epsilon), \tag{62}$$

and consequently

$$\Delta \le \tau(\rho,\epsilon). \tag{63}$$

Hence the generalization penalty and convergence perturbation in Theorems 6.2 and 6.3 are bounded in terms of (ρ,ϵ).

Proof. By construction, channels with small $\mathcal{I}_c^{(\ell)}$ (product of NDI and normalized weight norm) are removed preferentially. Let $w_c := \bar{w}_c^{(\ell)}$ denote the normalized weight score and $g_c := \mathrm{NDI}_c^{(\ell)}$. Sorting channels by $\mathcal{I}_c = g_c w_c$ and removing the lowest ρ fraction means that most removed channels have either small g_c or small w_c (or both).

Formally, consider the joint empirical measure of pairs (g_c, w_c) over all channels. Define level sets

$$S_\alpha := \{(c,\ell) \,:\, g_c \ge \alpha\}, \qquad \alpha \in [0,1]. \tag{64}$$

If the retained set has average NDI at least $1-\epsilon$, then a large mass of channels satisfies $g_c \ge 1-\epsilon'$ for small $\epsilon' \le O(\epsilon)$. The channels removed must therefore largely lie in the complement $\{g_c < 1-\epsilon'\}$. On that complement, by the ordering property of $\mathcal{I}_c = g_c w_c$, the cumulative normalized weight mass of removed channels is bounded above by the cumulative weight mass of channels with small g_c. Concretely, using Markov/Chebyshev style tail bounds over the empirical distribution, one can show

$$\sum_{(c,\ell)\in\mathcal{P}} w_c \le \frac{\epsilon'}{1-\epsilon'} \cdot \sum_{(c,\ell)} w_c \tag{65}$$

for an appropriately chosen ϵ' depending on ϵ and ρ. Rescaling back to Frobenius norms via

$$\|W_c^{(\ell)}\|_F \approx w_c \cdot \Big(\frac{1}{C_\ell}\sum_{c'} \|W_{c'}^{(\ell)}\|_F + \epsilon_w\Big), \tag{66}$$

we obtain an upper bound of the form $\tau(\rho,\epsilon)$. The exact functional form of τ depends on the empirical CDFs of g_c and w_c, but it is monotone increasing in ρ and in ϵ. This completes the constructive sketch. □

6.5. Matrix Concentration Bounds for Covariance / Correlation Estimation

We derive high-probability operator-norm bounds for the sample covariance and show how these translate into bounds for the sample correlation matrix used in the NDI spectral component.

Assumption 6.5 (Sub-Gaussian channel activations)**.** Let $\{z^{(t)}\}_{t=1}^N \subset \mathbb{R}^C$ be independent mean-zero activation vectors for a fixed layer (after mean subtraction) with coordinates corresponding to channels. There exists $\sigma > 0$ such that for every unit vector $u \in \mathbb{R}^C$ and every t,

$$\mathbb{E}\big[\exp(\lambda\, u^\top z^{(t)})\big] \le \exp\big(\tfrac{\lambda^2\sigma^2}{2}\big), \quad \forall \lambda \in \mathbb{R}. \tag{67}$$

Define the population covariance $\Sigma := \mathbb{E}[z^{(t)} z^{(t)\top}]$ and the sample covariance

$$\widehat{\Sigma} = \frac{1}{N}\sum_{t=1}^{N} z^{(t)} z^{(t)\top}. \tag{68}$$

Theorem 6.6 (Matrix Bernstein bound for covariance)**.** *Under Assumption 6.5 there exist universal constants $c_1, c_2 > 0$ such that for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\widehat{\Sigma} - \Sigma\|_2 \le c_1\sigma^2 \left(\sqrt{\frac{\log(2C/\delta)}{N}} + \frac{\log(2C/\delta)}{N}\right). \tag{69}$$

Proof. This is a standard consequence of matrix concentration inequalities (matrix Bernstein / Tropp). Treat each summand $X_t = z^{(t)} z^{(t)\top} - \Sigma$, note $\mathbb{E}[X_t] = 0$ and that the sub-Gaussian condition implies a uniform bound on the mgf of quadratic forms and a bound on $\|X_t\|_2$ with high probability. Applying matrix Bernstein (see Tropp-type bounds) yields the stated operator norm rate; the two-term expression reflects the usual variance and tail terms. □

From covariance to correlation. The NDI uses a *correlation matrix* $R = D^{-1/2}\Sigma D^{-1/2}$ where $D = \mathrm{diag}(\Sigma)$ is the diagonal of Σ (variances per channel). Let $\widehat{D} = \mathrm{diag}(\widehat{\Sigma})$ and $\widehat{R} = \widehat{D}^{-1/2}\widehat{\Sigma}\widehat{D}^{-1/2}$. We want a high-probability bound on $\|\widehat{R} - R\|_2$.

Proposition 6.7 (Correlation perturbation bound)**.** *Assume Assumption 6.5 and let $\lambda_{\min}^D := \min_c \Sigma_{cc} > 0$. Then with probability at least $1-\delta$,*

$$\|\widehat{R} - R\|_2 \le \frac{C'}{\lambda_{\min}^D}\,\|\widehat{\Sigma} - \Sigma\|_2, \tag{70}$$

where C' is a constant that depends on $\lambda_{\min}^D$ and on upper bounds for $\|\Sigma\|_2$ (explicit constants follow from standard perturbation expansions of $D^{-1/2}$).

Proof. Write

$$\widehat{R}-R=\widehat{D}^{-1/2}(\widehat{\Sigma}-\Sigma)\widehat{D}^{-1/2} \\ +(\widehat{D}^{-1/2}-D^{-1/2})\Sigma\widehat{D}^{-1/2} \\ +D^{-1/2}\Sigma(\widehat{D}^{-1/2}-D^{-1/2}). \tag{71}$$

Each term is bounded by products of $\|\widehat{\Sigma}-\Sigma\|_2$ and $\|\widehat{D}^{-1/2}-D^{-1/2}\|_2$. The latter can be bounded by a Lipschitz-type inequality for the map $x\mapsto x^{-1/2}$ on the positive reals and depends inversely on λ^D_{min}. Collecting terms yields the stated linear dependence on $\|\widehat{\Sigma}-\Sigma\|_2$ up to constants. □

Explicit rate combined. Combining Theorem 6.6 and Proposition 6.7 gives that with probability $1-\delta$,

$$\|\widehat{R}-R\|_2\le C''\sigma^2\left(\sqrt{\frac{\log(2C/\delta)}{N}}+\frac{\log(2C/\delta)}{N}\right), \tag{72}$$

for some constant C'' depending on λ^D_{min} and $\|\Sigma\|_2$. In particular, for $N\gtrsim\log C$ the dominant term is $O\big(\sqrt{\log C/N}\big)$.

6.6. Asymptotic Consistency of NDI

We now show that, under natural regularity conditions and as $N\to\infty$, each component of NDI_c (spectral diversity d_c, entropy informativeness u_c, and Hessian-based sensitivity $\tilde{s}_c$) converges to its population counterpart; hence the estimated NDI converges.

Assumption 6.8 (Regularity for entropy estimation)**.** Channel activations have a density with respect to Lebesgue measure (or are continuous) and have uniformly bounded support or bounded moments up to some order. The number of histogram/quantile bins $B=B(N)$ grows slowly with N, e.g. $B(N)\to\infty$ and $B(N)/N\to 0$.

Assumption 6.9 (Hessian probe consistency)**.** The empirical Hessian diagonal estimates via m Rademacher probes are unbiased estimates of the population Hessian diagonal on the representative mini-batch, and the variance of the Hutchinson estimator decays as $O(1/m)$. The representative mini-batch size used for Pearlmutter Hv computations grows with N or is sufficiently large to control sampling error.

Theorem 6.10 (Consistency of NDI components)**.** *Under Assumptions 6.5, 6.8, and 6.9, and if $B(N)$ and $m(N)$ satisfy $B(N)/N\to 0$ and $m(N)\to\infty$ slowly, then for each fixed channel c,*

$$d_c^{(N)}\xrightarrow{P}d_c,\qquad u_c^{(N)}\xrightarrow{P}u_c,\qquad \tilde{s}_c^{(N)}\xrightarrow{P}\tilde{s}_c, \tag{73}$$

where the right-hand side quantities are population values defined analogously but with expectations and population covariance/Hessian. Consequently,

$$\text{NDI}_c^{(N)}\xrightarrow{P}\text{NDI}_c \tag{74}$$

(pointwise convergence in probability).

Proof. Spectral diversity. By Theorem 6.6 and Proposition 6.7, $\|\widehat{R}-R\|_2\xrightarrow{P}0$. Standard eigenvalue/eigenvector perturbation results (Weyl + Davis–Kahan) then imply that the empirical eigenvalues and eigenvectors converge to the population ones, and hence the per-channel loadings $a_{c,i}^{(N)}$ and derived quantities $\phi_c^{(N)}$ converge in probability to their population counterparts. After min–max normalization (continuous map, provided denominators stay bounded away from zero, which holds w.h.p.), $d_c^{(N)}\to d_c$.

Entropy. Under Assumption 6.8, the histogram/quantile plug-in entropy estimator with Laplace smoothing is consistent provided $B(N)\to\infty$ slowly and $B(N)/N\to 0$ (standard density/entropy estimation theory). The bias term $(B-1)/(2N)$ used already is $o(1)$ under $B(N)=o(N)$. Hence $u_c^{(N)}\to u_c$.

Hessian diagonal via Hutchinson. Under Assumption 6.9, the Hutchinson estimator for the diagonal is unbiased and its variance decays as $O(1/m)$. With $m(N)\to\infty$ we obtain consistency for per-parameter diagonal entries; aggregating per-channel (finite sums) preserves consistency. Min–max normalization is continuous and so $\tilde{s}_c^{(N)}\to\tilde{s}_c$.

NDI multiplicative coupling. The mapping

$$(d,u,\tilde{s})\mapsto(d+\epsilon_f)^p(u+\epsilon_f)^q(\tilde{s}+\epsilon_f)^r \tag{75}$$

is continuous; hence convergence of the three components implies convergence of the product. This yields pointwise consistency of $\text{NDI}_c^{(N)}$. □

Uniform convergence remark. With stronger conditions (uniform sub-Gaussian tails across channels, moment bounds, and control of C relative to N) one can strengthen pointwise convergence to uniform convergence over channels (i.e., $\sup_c|\text{NDI}_c^{(N)}-\text{NDI}_c|\to 0$ in probability), which is useful for global ranking stability. This requires using matrix concentration with explicit $\log C$ dependence (as in Theorem 6.6) and uniform entropy estimation bounds.

6.7. Davis-Kahan, and Block Perturbation

We present perturbation results for eigenspaces and projectors. These sharpen the additive term $\frac{2\delta}{\gamma}$ appearing in the incoherence bound and clarify constant dependence.

Theorem 6.11 (Eigenvalue and eigenspace perturbation)**.** *Let R and $\widehat{R}$ be symmetric with $\widehat{R}=R+E$ and $\|E\|_2=\delta$. Let $\lambda_1\ge\cdots\ge\lambda_C$ be eigenvalues of R and $\widehat{\lambda}_i$ those of $\widehat{R}$. Fix k and assume $\gamma:=\lambda_k-\lambda_{k+1}>0$ and $\delta<\gamma/2$. Then:*

(a) (Weyl) For every i,

$$|\widehat{\lambda}_i - \lambda_i| \le \delta. \tag{76}$$

(b) (Davis-Kahan) If P_k and $\widehat{P}_k$ denote top-k projectors,

$$\|\sin\Theta(P_k, \widehat{P}_k)\|_2 \le \frac{\|E\|_2}{\gamma - \|E\|_2} \le \frac{\delta}{\gamma - \delta}. \tag{77}$$

(c) Consequently,

$$\|P_k - \widehat{P}_k\|_2 \le 2\,\frac{\delta}{\gamma - \delta}. \tag{78}$$

Proof. (a) is Weyl's inequality. (b) follows from the Davis–Kahan $\sin\Theta$ theorem in its refined form where the denominator uses the *separation* between spectral clusters: $\mathrm{sep}(\Lambda_1, \Lambda_2) \ge \gamma - \|E\|_2$ when the perturbation is small. The bound for $\|P_k - \widehat{P}_k\|_2$ then follows from standard relationships between $\sin\Theta$ and projector difference (a factor of 2 arises from triangle/identity decompositions). □

Improved incoherence bound. The proof of Lemma 6.1 gives the improved additive perturbation term:

$$\big|\langle z_c, z_j\rangle\big| \le \sqrt{\mu_c\mu_j} + \frac{4\delta}{\gamma - \delta}, \tag{79}$$

valid whenever $\delta < \gamma/2$ (so denominator remains $> \gamma/2$ and the bound is $O(\delta/\gamma)$ with better constant control).

6.8. Explicit Sample Complexity for Stable NDI Ranking

We conclude by combining the concentration and perturbation results to give an explicit sampling requirement such that the eigenspace perturbation term in the incoherence bound is below a target $\eta > 0$.

Corollary 6.12 (Sample complexity for controlled projector perturbation)**.** *Under Assumption 6.5 and the notation above, fix a desired projector error tolerance $\varepsilon \in (0, \gamma/4)$. There exist constants $C_1, C_2 > 0$ such that if*

$$N \ge C_1\sigma^4 \frac{\log(C/\delta)}{\varepsilon^2}, \tag{80}$$

then with probability at least $1 - \delta$ we have $\|\widehat{R} - R\|_2 \le \varepsilon$ and hence

$$\|P_k - \widehat{P}_k\|_2 \le 2\frac{\varepsilon}{\gamma - \varepsilon} \le \frac{4\varepsilon}{\gamma}, \tag{81}$$

so the additive term in the incoherence bound is at most $O(\varepsilon/\gamma)$. Concretely, choosing $\varepsilon = \eta\gamma/4$ yields the projector perturbation $\le \eta$.

Proof. Solve the inequality in Theorem 6.6 for N to make the RHS $\le \varepsilon$ (dominant term scales as $\sigma^2\sqrt{\log C/N}$). The stated N suffices. □

6.9. Algorithm: Neural Differentiation Pruning (NDP)

The overall procedure is summarized in Algorithm 1, which details the computation of NDI, its integration with normalized weight norms, and the global ranking-based pruning process.

Algorithm 1 Neural Differentiation Pruning (NDP)

```
Require: pre-trained parameters $\Theta$, target sparsity $\rho$, batch count $T$, bins $B$, Hessian approximation parameters
Collect mean-pooled activations $z_c^{(\ell)}$ over $T$ mini-batches
for each layer $\ell$ do
    Compute normalized covariance $R^{(\ell)}$ with shrinkage
    Eigendecompose $R^{(\ell)} = V\Lambda V^\top$ (or randomized SVD)
    Compute diversity $d_c^{(\ell)}$, informativeness $u_c^{(\ell)}$, and sensitivity $\tilde{s}_c^{(\ell)}$
    Compute $\mathrm{NDI}_c^{(\ell)} = \big(d_c^{(\ell)} + \epsilon_f\big)^p \cdot \big(u_c^{(\ell)} + \epsilon_f\big)^q \cdot \big(\tilde{s}_c^{(\ell)} + \epsilon_f\big)^r$
end for
for each channel $c$ do
    Compute weight norm $\|W_c^{(\ell)}\|_F$
    Compute normalized weight $\bar{w}_c^{(\ell)} = \frac{\|W_c^{(\ell)}\|_F}{\frac{1}{C_\ell}\sum_{c'}\|W_{c'}^{(\ell)}\|_F + \epsilon_w}$
    Compute importance $\mathcal{I}_c^{(\ell)} = \mathrm{NDI}_c^{(\ell)} \cdot \bar{w}_c^{(\ell)}$
end for
Sort all channels by $\mathcal{I}_c^{(\ell)}$ and prune the lowest $\rho$ fraction globally
Fine-tune pruned model
```

6.10. More Experiment Results

6.10.1. Experiments on MLP-Net

We further evaluate NDP on a fully connected MLP-Net trained on MNIST. Figure 3 left reports test accuracy under increasing weight sparsity, comparing NDP with several representative pruning methods. Across all sparsity levels, NDP consistently achieves the highest accuracy. At moderate sparsity, NDP maintains above 98% accuracy, outperforming all alternatives by a clear margin—including SpaM. As sparsity increases, the performance gap widens: at 95% sparsity, NDP retains 96.68% accuracy, whereas MSP and SpaM drop to 94.70% and 89.43%, respectively. Under extreme sparsity, NDP still preserves 94.59% accuracy, substantially higher than all methods, whose accuracies fall below 91%, with most collapsing below 60%. These results demonstrate that pruning using NDI leads to significantly improved resilience to aggressive sparsifica-

Table 6. For ResNet-18 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches, including Cropit, EarlyCrop and EarlySNAP [45], SNAP [54]. **Neural sparsity (%)**.

	50	60	70	75	80	85	90
CroPit-S	92.15	91.18	90.98	90.00	88.90	88.10	85.10
EarlyCroP-S	92.33	92.23	91.75	91.35	90.78	87.85	84.18
EarlySNAP	92.08	92.33	92.00	91.25	90.43	88.60	83.50
SNAP	91.93	91.48	91.23	90.48	89.40	87.55	85.45
NDP	**94.00**	**93.96**	**93.48**	**92.96**	**92.60**	**91.28**	**89.94**

Table 7. For ResNet-18 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches, including Cropit, EarlyCrop and EarlySNAP [45], SNAP [54]. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
CroPit-S	92.70	92.26	91.66	91.06	90.41	89.22	88.58
EarlyCroP-S	92.55	92.40	92.31	92.19	91.31	90.52	88.57
EarlySNAP	92.40	92.20	92.13	92.18	91.24	90.36	89.01
SNAP	92.19	91.96	91.74	91.38	90.80	89.89	88.83
NDP	**94.36**	**94.14**	**93.96**	**93.41**	**92.56**	**91.20**	**90.03**

tion, enabling compression rates at which existing methods experience severe degradation.

6.10.2. Experiments on CIFAR-10

Tables 6 and 7 (Figure 4) present a detailed comparison of pruning methods on ResNet-18 trained on the CIFAR-10 dataset. Similarly, Tables 8 and 9 (Figure 5) provide results for VGG-16. Across both network architectures, our proposed NDP method consistently outperforms existing pruning techniques—including CroPit, EarlyCrop, EarlySNAP, and SNAP—under both neural and weight sparsity settings. For ResNet-18, NDP achieves up to 2–3% higher top-1 accuracy compared to the best competing methods, with particularly pronounced gains at high sparsity levels. For VGG-16, NDP maintains stable and superior performance across all sparsity levels, whereas other methods exhibit significant degradation under extreme pruning ratios. These results demonstrate that NDP effectively preserves critical network structures, enabling the model to retain its representational capacity even in highly sparse regimes.

Table 8. For VGG-16 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches. **Neural sparsity (%)**.

	50	60	70	75	80	85	90
CroPit-S	92.22	92.50	92.25	92.22	92.00	91.81	90.89
EarlyCroP-S	89.53	91.77	92.22	91.94	91.81	91.80	90.83
EarlySNAP	89.58	91.81	92.28	92.22	92.00	91.66	77.50
SNAP	91.39	92.50	92.08	92.22	91.94	91.39	86.53
NDP	**93.15**	**93.10**	**93.05**	**92.86**	**92.79**	**92.65**	**92.58**

Table 9. For VGG-16 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
CroPit-S	92.70	92.60	92.20	91.80	91.50	91.10	90.50
EarlyCroP-S	92.40	92.20	92.00	91.80	91.30	91.00	90.70
EarlySNAP	92.44	92.40	92.20	91.80	91.40	91.60	71.40
SNAP	92.10	92.00	92.20	91.80	90.90	87.30	78.20
NDP	**93.28**	**93.22**	**93.16**	**93.05**	**93.03**	**92.93**	**92.81**

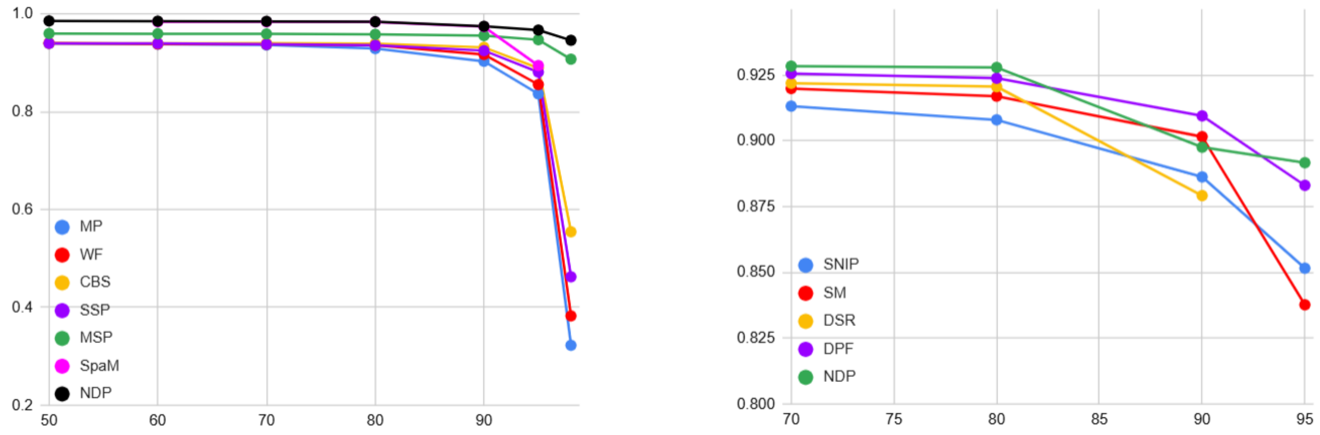
Table 10. Top-1 test accuracy on CIFAR-10 for weight pruning.

Model	Methods					Sparsity
	SNIP ([31])	SM	DSR([38])	DPF([52])	**NDP**	
ResNet-20	91.32	91.99	92.19	92.56	**92.84**	70%
	90.80	91.70	92.06	92.38	**92.79**	80%
	88.63	90.16	87.92	90.95	**89.76**	90%
	85.16	83.77	*	88.31	**89.17**	95%
ResNet-32	90.66	91.72	91.64	92.60	**93.33**	90%
	87.52	88.90	84.44	91.29	**92.05**	95%
ResNet-56	91.77	92.94	93.98	94.06	**94.68**	90%
	*	91.36	92.66	92.82	**94.32**	95%

Beyond a single architecture, Table 10 (Figure 3 right and 6) evaluates NDP on multiple ResNet variants, comparing against SOTA pruning methods. Across a wide range of sparsity levels, NDP consistently achieves the highest accuracy. Notably, the performance gap widens as the sparsity level increases, highlighting the robustness of NDP in extreme pruning regimes. Furthermore, the benefits of NDP are amplified with increasing model depth: on ResNet-56, NDP delivers up to 2% higher accuracy than the next-best method at 95% sparsity. These findings indicate that incorporating NDI through our proposed criterion allows NDP to adaptively preserve essential neurons and weights, offering strong generalization and stability under aggressive compression. This makes NDP a promising choice for deploying efficient yet accurate models in resource-constrained environments.

6.10.3. Experiment on DenseNet-121

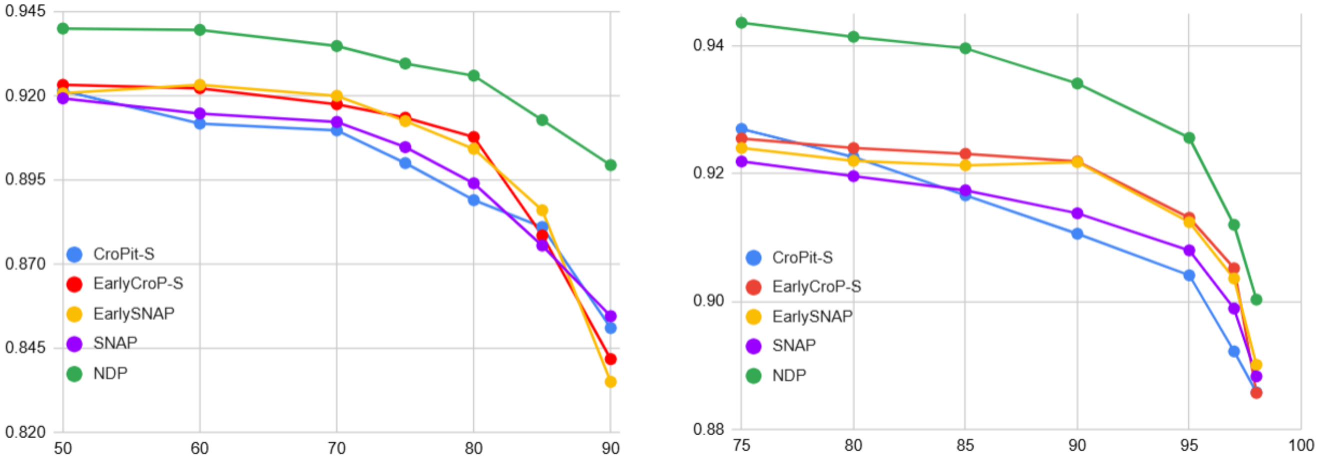
We further validate the effectiveness of NDP on DenseNet-121 trained on CIFAR-10, comparing against a range of SOTA pruning methods. Table 11 (Figure 7 left) reports performance under aggressive weight sparsity ratios of 95.5% and 98.85%. Classical pruning strategies such as Global pruning and E-R ker. pruning exhibit severe degradation at extreme sparsity, with accuracies dropping below 60%. More adaptive approaches such as LAMP, SuRP, and RDP substantially alleviate this collapse, yet still suffer non-trivial accuracy losses when sparsity exceeds 95%. In



(a) MLP-Net on MNIST

(b) ResNet-20 on CIFAR-10

Figure 3. NDP also outperforms other approaches for MLP-Net and ResNet-20 networks trained on MNIST and CIFAR-10. **Left**: MLP-Net on MNIST. **Right**: ResNet-20 on CIFAR-10.



(a) Neuron Sparsity

(b) Weight Sparsity

Figure 4. For ResNet-18 networks on CIFAR-10, NDP can find sparser solutions maintaining better performance than other approaches. **Left**: Neural sparsity. **Right**: Weight sparsity.

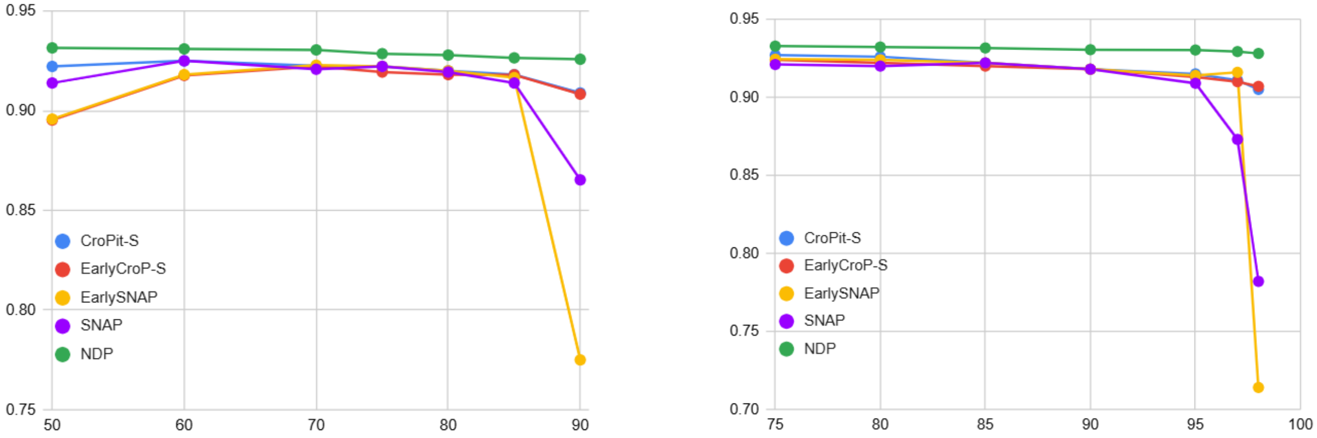
contrast, NDP consistently delivers superior performance, achieving 93.15% at 95.5% sparsity and 91.83% at 98.85%, significantly outperforming the strongest method. The gap widens at ultra-high sparsity, highlighting NDP's capacity to preserve discriminative features even when the parameter budget is extremely constrained. These results confirm that NDP at the neuron level not only mitigates the risk of representational collapse but also scales more robustly to dense connectivity patterns such as those in DenseNet architectures.

6.10.4. Experiment on Tiny-ImageNet

To further validate the effectiveness of NDP, we evaluate its performance on the challenging Tiny-ImageNet dataset using the ResNet-18. Figure 8 reports the comparison against representative SOTA pruning methods, including SNIP, Iterative-SNIP, SynFlow, PHEW, and NBP, across a wide range of sparsity levels from moderate to extreme. The results demonstrate that NDP consistently outperforms all methods by a substantial margin in terms of top-1 clas-

Table 11. For DenseNet-121 on CIFAR-10. NDP again outperforms the other pruning approachs. **Weight sparsity (%)**.

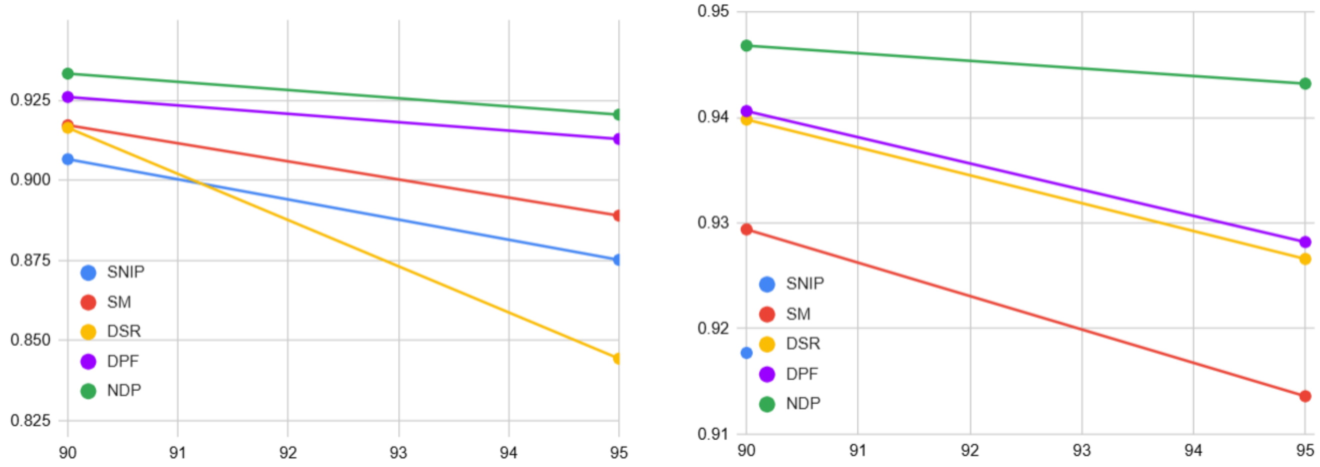
	95.5	98.85
Global([37])	*	45.30 ± 27.75
E-R ker.([13])	*	59.06 ± 25.61
LAMP([22])	90.11 ± 0.13	85.13 ± 0.31
SuRP([21])	90.75	86.71
RDP([58])	91.49 ± 0.21	87.70 ± 0.24
Our NDP	**93.15 ± 0.23**	**91.83 ± 0.18**



(a) Neuron Sparsity

(b) Weight Sparsity

Figure 5. The results of VGG-16 on CIFAR-10. NDP better maintains performance at higher sparsities than other approaches. **Left**: Neural sparsity. **Right**: Weight sparsity.



(a) ResNet-20 on CIFAR-10

(b) ResNet-32 on CIFAR-10

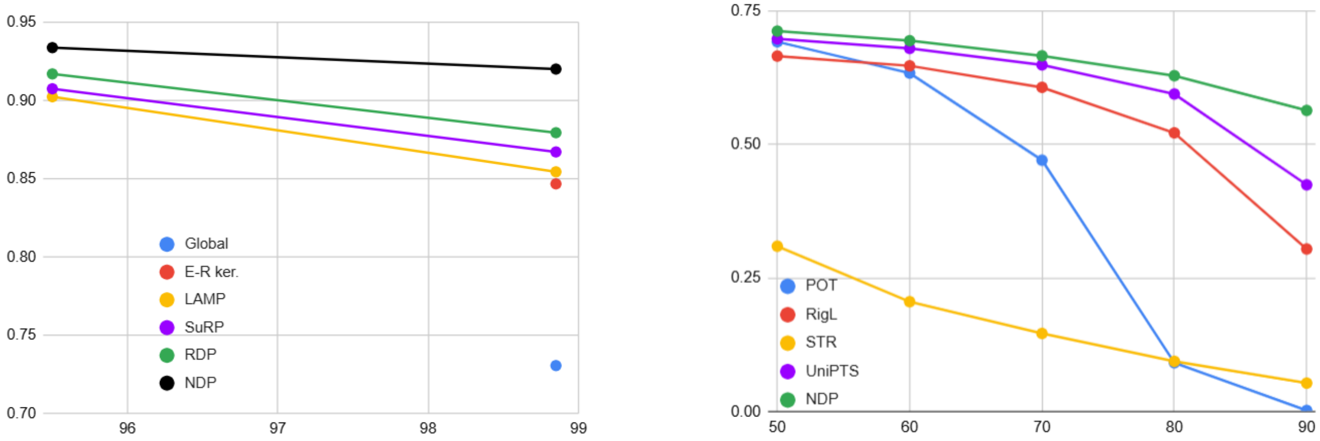
Figure 6. NDP also outperforms other approaches for ResNet-32 and ResNet-56 networks trained on CIFAR-10. **Left**: ResNet-32 on CIFAR-10. **Right**: ResNet-56 on CIFAR-10.

sification accuracy while simultaneously reducing computational cost measured in FLOPs. At moderate sparsity, NDP achieves 72.10% accuracy, exceeding the second-best method by nearly 14 percentage points. This advantage becomes even more pronounced as the sparsity level increases. For example, under 96.84% sparsity, NDP maintains 60.23% accuracy, which is over 9 percentage points higher than PHEW and nearly 11 points higher than SynFlow. At the extreme pruning regime, NDP achieves 53.63% accuracy, whereas all other methods collapse below 41.05%. In addition to accuracy, NDP exhibits a highly favorable FLOPs-accuracy trade-off. For instance, at 90% sparsity, NDP reduces FLOPs to 2.32×10^8, which is less than half of PHEW and SynFlow, while simultaneously achieving 66.32% accuracy compared to 55.93% and 54.68%, respectively. Even at extreme sparsity, NDP preserves strong accuracy with an ultra-lightweight computational budget of only 0.28×10^8 FLOPs.

These results confirm that NDP not only retains significantly more discriminative power than existing pruning approaches but also achieves superior efficiency. The consistent dominance across varying sparsity regimes highlights the robustness of NDP. This demonstrates the key advantage of incorporating NDI as a principled criterion for pruning, allowing the network to selectively retain highly distinctive neurons while aggressively eliminating redundant ones.

6.10.5. Experiment on ImageNet

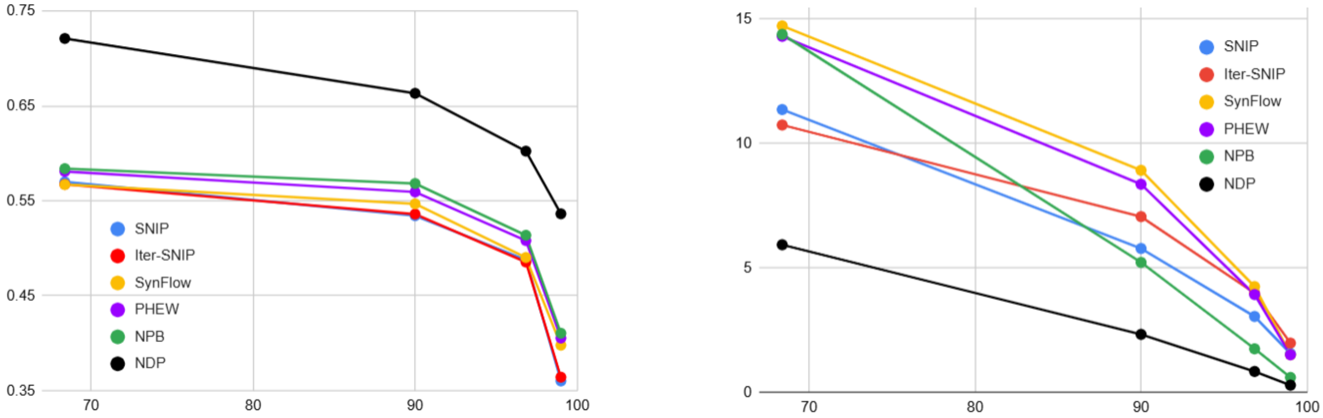
We evaluate the proposed NDP method on the ImageNet dataset using the MobileNet-V2 and compare it against



(a) DenseNet-121 on CIFAR-10

(b) MobileNet-V2 on ImageNet

Figure 7. NDP also outperforms other approaches for DenseNet-121 on CIFAR-10 and MobileNet-V2 on ImageNet. **Left**: DenseNet-121 on CIFAR-10. **Right**: MobileNet-V2 on ImageNet.



(a) Weight Sparsity

(b) FLOPs (10^8)

Figure 8. NDP also outperforms other approaches for ResNet-18 networks trained on Tiny-ImageNet. **Left**: Weight sparsity. **Right**: FLOPs(10^8).

several SOTA pruning approaches, including POT, RigL, STR, and UniPTS. Figure 7 right reports the Top-1 accuracy across different pruning ratios ranging from 50% to 90%. As shown, NDP consistently outperforms all methods under every sparsity level. In particular, at moderate pruning levels, NDP achieves 69.45% and 66.62% Top-1 accuracy, surpassing the next best method, UniPTS, by margins of 1.44% and 1.69%, respectively. Even at extreme sparsity, NDP maintains a strong 56.39% accuracy, whereas other methods suffer significant degradation. These results demonstrate that NDP yields superior robustness to aggressive pruning while preserving competitive performance on large-scale datasets.

6.10.6. Leaky ReLU

Our pruning framework is fundamentally motivated by the principle of Neural Differentiation, which emphasizes that each neuron should contribute distinct representational information to the network. Neurons that fail to differentiate from others—exhibiting redundant or consistently uninformative activation patterns—can be pruned without impairing model expressivity. While this mechanism is naturally pronounced in ReLU activations, where neurons can become completely inactive due to the zeroing effect on negative inputs, the situation is less clear for Leaky ReLU [48, 50]. The small negative slope in Leaky ReLU prevents absolute inactivity, but neurons predominantly confined to the negative activation regime still provide little discriminative capacity and may be regarded as functionally redundant.

Table 12. ResNet-18 networks with *Leaky ReLU* trained on CIFAR-10. NDP again outperforms the other pruning methods. **Neural sparsity (%)**.

	50	60	70	75	80	85	90
EarlyCroP-S	88.89	88.97	88.17	87.10	85.99	84.72	80.71
EarlySNAP	89.40	87.86	87.02	85.99	85.00	84.33	80.00
SNAP	88.21	87.62	86.67	86.07	85.48	81.75	78.33
NDP	**94.04**	**93.84**	**92.60**	**92.33**	**91.64**	**91.03**	**89.65**

Table 13. ResNet-18 networks with *Leaky ReLU* trained on CIFAR-10. NDP again outperforms the other pruning methods. **Weight sparsity (%)**.

	75	80	85	90	95	97	98
EarlyCroP-S	88.95	88.90	88.94	88.78	87.56	86.34	85.48
EarlySNAP	89.33	89.39	88.78	87.72	86.66	85.59	84.78
SNAP	88.34	88.22	87.85	87.29	86.27	85.39	83.32
NDP	**93.90**	**93.77**	**93.28**	**92.82**	**91.98**	**89.87**	**88.23**

To evaluate this hypothesis, we applied NDP to ResNet-18 trained on CIFAR-10 with Leaky ReLU activations. NDP identifies neurons with low differentiation power—those whose activations remain clustered within uninformative subspaces—and eliminates them to encourage a more diverse representational basis. Tables 12 and 13 (Figure 9) report the performance under varying levels of neural sparsity and weight sparsity, respectively. The results demonstrate that, NDP consistently outperforms strong methods such as EarlyCroP-S, EarlySNAP, and SNAP. Under neural sparsity constraints, NDP achieves accuracies above 92% even at 75–80% sparsity, whereas competing methods degrade more sharply, dropping below 86%. Similarly, under weight sparsity constraints, NDP preserves predictive performance above 91% at 95% sparsity, while other approaches fall below 87%. These results highlight that pruning guided by NDI—rather than simple magnitude or early-activation heuristics—offers greater resilience against accuracy degradation. Collectively, these findings reinforce our central claim: NDP fosters robust representational diversity, enabling networks to maintain high accuracy even under extreme sparsification, and extending its effectiveness beyond standard ReLU activations to Leaky ReLU networks.

References

[1] Sotiris Anagnostidis, et al. Dynamic context pruning for efficient and interpretable autoregressive transformers. *Advances in Neural Information Processing Systems*, 36: 65202–65223, 2023. 2, 3

[2] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Advances in Neural Information Processing Systems*, 36:9707–9750, 2023. 2

[3] Thomas A Bos, et al. A systematic review and embryological perspective of pluripotent stem cell-derived autonomic postganglionic neuron differentiation for human disease modeling. *Elife*, 14:e103728, 2025. 3

[4] Wenyu Chen, et al. Network pruning at scale: A discrete optimization approach. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. 7

[5] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020. 2

[6] Minsik Cho, et al. Pdp: Parameter-free differentiable pruning is all you need. *Advances in Neural Information Processing Systems*, 36:45833–45855, 2023. 2

[7] Gabriele Ciceri, et al. An epigenetic barrier sets the timing of human neuronal maturation. *Nature*, 626(8000):881–890, 2024. 3

[8] Arthur Conmy, et al. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023. 2

[9] Pau de Jorge, et al. Progressive skeletonization: Trimming more fat from a network at initialization. *International Conference on Learning Representations*, 2021. 8

[10] Rayen Dhahri, et al. Shaving weights with occam's razor: Bayesian sparsification for neural networks using the marginal likelihood. *Advances in Neural Information Processing Systems*, 37:24959–24989, 2024. 7

[11] Shibhansh Dohare, et al. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024. 2

[12] Mohamed Elsayed and A Rupam Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[13] Utku Evci, et al. Rigging the lottery: Making all tickets winners. *International conference on machine learning*, page 2943–2952, 2020. 8, 7

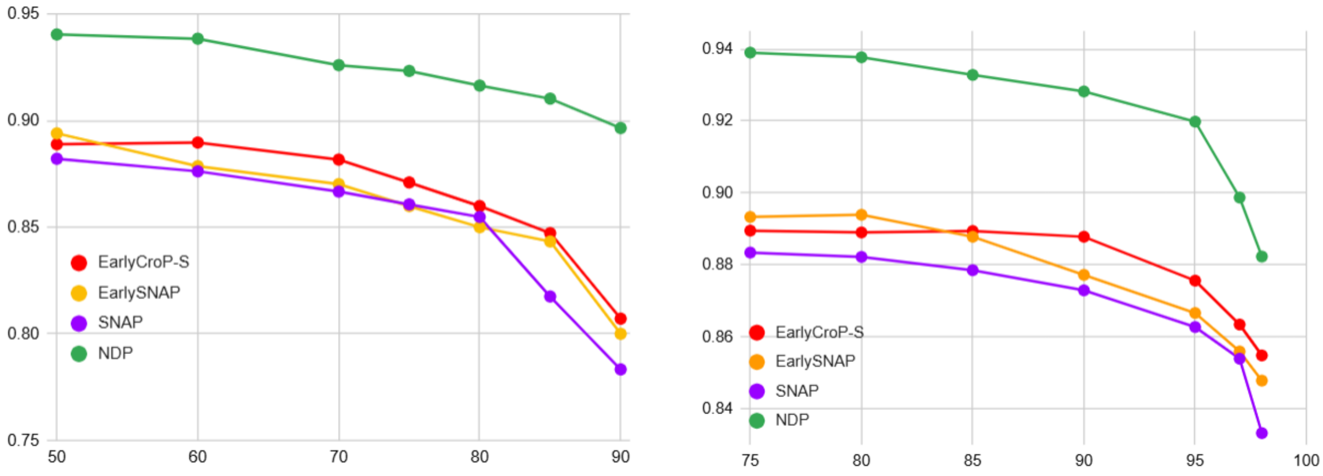
[14] Gongfan Fang, et al. Structural pruning for diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[15] Gongfan Fang, et al. Isomorphic pruning for vision models. In *European Conference on Computer Vision*, pages 232–250. Springer, 2024. 2

[16] Shangqian Gao, et al. Bilevelpruning: unified dynamic and static channel pruning for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16090–16100, 2024. 2, 3

[17] Erik Gärtner, et al. Transformer-based learned optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11970–11979, 2023. 2

[18] Fang Gongfan, et al. Depgraph: Towards any structural pruning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 16091–16101, 2023. 2



(a) Neuron Sparsity

(b) Weight Sparsity

Figure 9. ResNet-18 networks with *Leaky ReLU* trained on CIFAR-10. NDP again outperforms the other pruning methods. **Left**: Neural sparsity. **Right**: Weight sparsity.

[19] Sayed Hatefi, et al. Pruning by explaining revisited: Optimizing attribution methods to prune cnns and transformers. In *European Conference on Computer Vision*, pages 152–169. Springer, 2024. 2

[20] Yaomin Huang, et al. Cp3: Channel pruning plug-in for point-based networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5302–5312, 2023. 2

[21] Berivan Isik, et al. An information-theoretic justification for model pruning. In *International Conference on Artificial Intelligence and Statistics*, pages 3821–3846. PMLR, 2022. 7

[22] Lee Jaeho, et al. Layer-adaptive sparsity for the magnitude-based pruning. In *International Conference on Learning Representations*, 2021. 7

[23] Jinghan Jia, et al. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023. 2

[24] Agustinus Kristiadi, et al. The geometry of neural nets' parameter spaces under reparametrization. *Advances in Neural Information Processing Systems*, 36:17669–17688, 2023. 2

[25] Eldar Kurtić, et al. Ziplm: Inference-aware structured pruning of language models. *Advances in Neural Information Processing Systems*, 36:65597–65617, 2023. 2

[26] Aditya Kusupati, et al. Soft threshold weight reparameterization for learnable sparsity. In *International Conference on Machine Learning*, pages 5544–5555. PMLR, 2020. 8

[27] Andrey Kuzmin, et al. Pruning vs quantization: Which is better? *Advances in neural information processing systems*, 36:62414–62427, 2023. 2

[28] Denis Kuznedelev, et al. Cap: Correlation-aware pruning for highly-accurate sparse vision models. *Advances in Neural Information Processing Systems*, 36:28805–28831, 2023. 2

[29] Thomas Laurent, et al. Feature collapse. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[30] Ivan Lazarevich, et al. Post-training deep neural network pruning via layer-wise calibration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 798–805, 2021. 8

[31] Namhoon Lee, et al. Snip: Single-shot network pruning based on connection sensitivity. *International Conference on Learning Representations*, 2019. 8, 6

[32] Weihao Lin, et al. S2hpruner: Soft-to-hard distillation bridges the discretization gap in pruning. *Advances in Neural Information Processing Systems*, 37:116547–116574, 2024. 2, 3

[33] Shreyas Malakarjun Patil, et al. Phew: Constructing sparse networks that learn fast and generalize well without training data. *Proceedings of the 38th International Conference on Machine Learning*, 139:8432–8442, 2021. 8

[34] Pierre Marion and Raphaël Berthier. Leveraging the two-timescale regime to demonstrate convergence of neural networks. *Advances in Neural Information Processing Systems*, 36:64996–65029, 2023. 2

[35] Christian HX Mehmeti-Göpel and Michael Wand. On the weight dynamics of deep normalized networks. In *Forty-first International Conference on Machine Learning*, 2024. 2

[36] Hancheng Min, et al. Early neuron alignment in two-layer reLU networks with small initialization. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[37] Ari Morcos, et al. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32, 2019. 7

[38] Hesham Mostafa and Xin Wang. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *International Conference on Machine Learning*, pages 4646–4655. PMLR, 2019. 6

[39] Michael C Mozer and Paul Smolensky. Using relevance to

reduce network size automatically. *Connection Science*, 1 (1):3–16, 1989. 7

[40] Saurav Muralidharan, et al. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102, 2024. 3

[41] Neel Nanda, et al. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[42] Hongjing Niu, et al. Rectifying shortcut learning through cellular differentiation in deep learning neurons. In *BMVC*, 2024. 3

[43] Kiho Park, et al. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. 2

[44] Hoang Pham, et al. Towards data-agnostic pruning at initialization: What makes a good sparse mask? *Advances in Neural Information Processing Systems*, 36:80044–80065, 2023. 2, 8

[45] John Rachwan, et al. Winning the lottery ahead of time: Efficient early network pruning. *Proceedings of the 39th International Conference on Machine Learning*, 162: 18293–18309, 2022. 7, 6

[46] Sidak Pal Singh and Dan Alistarh. Woodfisher: Efficient second-order approximation for neural network compression. *Advances in Neural Information Processing Systems*, 33:18098–18109, 2020. 7

[47] Anna Soligo, et al. Inducing, detecting and characterising neural modules: A pipeline for functional interpretability in reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. 2

[48] Haoyuan Sun, et al. Entropy-based activation function optimization: A method on searching better activation functions. In *The Thirteenth International Conference on Learning Representations*, 2025. 9

[49] Hidenori Tanaka, et al. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in Neural Information Processing Systems*, 33:6377–6389, 2020. 8

[50] Keke Tang, et al. Simplification is all you need against out-of-distribution overconfidence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5030–5040, 2025. 9

[51] Quan Tang, et al. Dynamic token pruning in plain vision transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 777–786, 2023. 3

[52] Lin Tao, et al. Dynamic model pruning with feedback. In *International Conference on Learning Representations*, 2020. 6

[53] Yuchuan Tian, et al. Towards higher ranks via adversarial weight pruning. *Advances in Neural Information Processing Systems*, 36:1189–1207, 2023. 2

[54] Stijn Verdenius, et al. Pruning via iterative ranking of sensitivity statistics. *CoRR*, 2020. 7, 6

[55] Zheng Wang, et al. The implicit bias of gradient descent toward collaboration between layers: A dynamic analysis of multilayer perceptions. *Advances in Neural Information Processing Systems*, 37:74868–74898, 2024. 2

[56] Jingyang Xiang, et al. Subp: Soft uniform block pruning for $1 \times n$ sparse cnns multithreading acceleration. *Advances in Neural Information Processing Systems*, 36:52033–52050, 2023. 2

[57] Jingjing Xie, et al. Unipts: A unified framework for proficient post-training sparsity. In *CVPR*, pages 5746–5755, 2024. 8

[58] Kaixin Xu, et al. Efficient joint optimization of layer-adaptive weight pruning in deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17447–17457, 2023. 7

[59] Changdi Yang, et al. Pruning parameterization with bi-level optimization for efficient semantic segmentation on the edge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15402–15412, 2023. 2

[60] Ke Yi, et al. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023. 2

[61] Xin Yu, et al. The combinatorial brain surgeon: Pruning weights that cancel one another in neural networks. In *International Conference on Machine Learning*, pages 25668–25683. PMLR, 2022. 7

[62] Zheng Zhan, et al. Exploring token pruning in vision state space models. *Advances in Neural Information Processing Systems*, 37:50952–50971, 2024. 3

[63] Yuhui Zhang, et al. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[64] Shaochen Zhong, et al. One less reason for filter pruning: Gaining free adversarial robustness with structured grouped kernel pruning. *Advances in neural information processing systems*, 36:62032–62061, 2023. 2