Calibration-Aware Relevance Estimation for Structured Filter Pruning

Anonymous authors
Paper under double-blind review

Abstract

Structured pruning is widely used to reduce the computational and memory requirements of deep neural networks (DNNs), enabling efficient deployment in resource-constrained computer vision systems. However, in safety-critical applications, predictive confidence must also be well calibrated, such that confidence accurately reflects the probability of correctness. Existing structured pruning methods primarily optimize predictive performance or feature importance, while largely overlooking uncertainty calibration (UC) as a pruning objective. To address this gap, we propose, to the best of our knowledge, the first calibration-aware sample selection strategy for relevance-based structured pruning. Instead of estimating filter importance over the entire validation set, the proposed framework identifies calibration-critical spillover samples from reliability diagrams and performs Layer-wise Relevance Propagation only on this informative subset, enabling pruning decisions that explicitly account for predictive reliability. We evaluate the proposed approach on convolutional neural networks (VGG-16, ResNet-34/56, and DenseNet-121) and a vision transformer (DeiT-Tiny) for the CIFAR-10, CIFAR-100, and ImageNet datasets. Experimental results demonstrate that the proposed method improves UC while maintaining competitive predictive accuracy under comparable sparsity levels, with reductions in expected calibration error (ECE) of over 50% relative to conventional structured pruning in the best-performing settings. Furthermore, the resulting pruned models remain amenable to post-hoc recalibration, reducing calibration error below 2% ECE in several pruning settings while preserving predictive performance. These findings establish UC not only as an evaluation metric but also as an effective optimization objective for structured model compression, improving the reliability of compressed deep neural networks without sacrificing deployment efficiency.

1 Introduction

Reliability issues and high resource consumption of computer vision deep neural networks (DNNs) limit their applicability to energy- and memory-constrained safety-relevant platforms (Han et al., 2016a). To address the resource demands of an already-trained DNN, post-training pruning techniques have been developed to reduce model complexity while preserving accuracy. In general, these methods fall into *unstructured pruning*, which removes individual weights and induces sparsity, and *structured pruning*, which removes entire architectural components such as convolutional filters, channels, neurons, or attention heads (Liu et al., 2017; Li et al., 2017a; Molchanov et al., 2016; Michel et al., 2019; Sajjad et al., 2023). Although unstructured pruning can achieve high sparsity, its irregular sparsity patterns typically require specialized sparse-computation support to yield practical acceleration on standard hardware (Wang, 2020). Structured pruning, by contrast, preserves dense computation patterns and is therefore generally more hardware-efficient, producing more reliable reductions in latency and memory footprint in deployment (Mao et al., 2017). In structured pruning, the resulting architecture depends not only on which components are pruned but also on the ranking criterion and the data used to evaluate it. Despite previous work on static and dynamic selection of prunable components (Park et al., 2023), pruning decisions remain primarily driven by accuracy-based objectives, which leaves the data subset used to estimate filter importance a critical, yet often overlooked, factor.

However, preserving accuracy alone is insufficient in safety-critical settings, where confidence estimates must remain reliable enough to support downstream decisions such as fallback triggering, uncertainty-aware control, or human intervention (Willers et al., 2020; Burton, 2023). In this context, a pruned model can remain accurate yet become poorly calibrated, producing overconfident or underconfident predictions that undermine its practical trustworthiness. Uncertainty calibration (UC) measures the alignment between predicted confidence and the empirical likelihood of correctness: for example, a perfectly calibrated model that predicts a class with 80% confidence should be correct 80% of the time. A standard tool for assessing UC is the reliability diagram, which groups predictions into confidence bins and compares average confidence with empirical accuracy within each bin. Miscalibrated models can produce overconfident outputs that bypass safety checks or underconfident outputs that delay decisions, compromising both safety and efficiency (Wang et al., 2023). Although post-hoc recalibration methods for trained DNNs exist (Platt, 1999; Guo et al., 2017), they offer limited correction and do not address the root causes of miscalibration.

Recent work suggests that substantial architectural changes induced by pruning can even improve UC (Mitra et al., 2024). In addition, alternative pruning objectives such as robustness (Sehwag et al., 2020) or interpretability (Yeom et al., 2021; Yu & Xiang, 2023; Voita et al., 2019; Ghadiri et al., 2024) have been shown to allow effective pruning without degrading predictive accuracy. This motivates an as-yet unexplored question: *can UC itself directly guide post-training pruning, so as to improve calibration without sacrificing the accuracy–compression tradeoff?* Although some pruning approaches use entropy-based criteria to mask prunable elements (Liao et al., 2023), and classical active learning methods rely on confidence-based measures for sample selection (Yang & Loog, 2016), no prior work has explicitly used UC as a guiding principle for pruning, neither through the pruning criterion itself nor through the selection of data used to estimate component importance. In this paper, we address this gap by presenting, to the best of our knowledge, the first calibration-aware sample selection scheme for relevance-based filter pruning that explicitly incorporates UC into post-training pruning with the goal of improving it.

In our framework, filter importance is estimated using Layer-wise Relevance Propagation (LRP; Yeom et al., 2021; Montavon et al., 2019; Achtibat et al., 2024), which attributes output behavior to internal network components and is applicable to both CNNs and vision transformers. Filters are ranked according to their aggregated relevance to the predicted class's confidence, irrespective of whether the prediction itself is correct, making LRP well suited for identifying structural components that contribute little to a model's confidence—and hence to its calibration. Unlike prior LRP-based pruning methods, which estimate filter importance using the *entire validation set*, we instead perform relevance estimation only on calibration-critical *spillover samples*: predictions that are incorrect, overconfident, or underconfident near calibration-bin boundaries. This rethinks sample selection for relevance-guided structured pruning by focusing filter importance estimation on the samples most informative for predictive calibration. Figure 1 provides an overview of the proposed framework.

Contributions. To the best of our knowledge, this is the first work to explicitly incorporate uncertainty calibration (UC) into structured filter pruning. Specifically, we:

1. propose a calibration-aware spillover sample selection strategy that identifies calibration-critical samples from reliability diagrams for relevance estimation;
2. develop a relevance-guided structured pruning framework that combines Taylor-based layer sensitivity analysis with LRP-based filter ranking computed exclusively on the selected spillover samples; and
3. demonstrate across CNNs and a vision transformer on CIFAR-10, CIFAR-100, and ImageNet that rethinking sample selection consistently improves uncertainty calibration while maintaining competitive predictive accuracy under comparable sparsity levels.

2 Related Work

Model Pruning. To reduce the computational cost of DNNs, *structured pruning* – and in particular *filter pruning* – has attracted considerable interest due to its strong empirical effectiveness (Liu et al., 2019) and

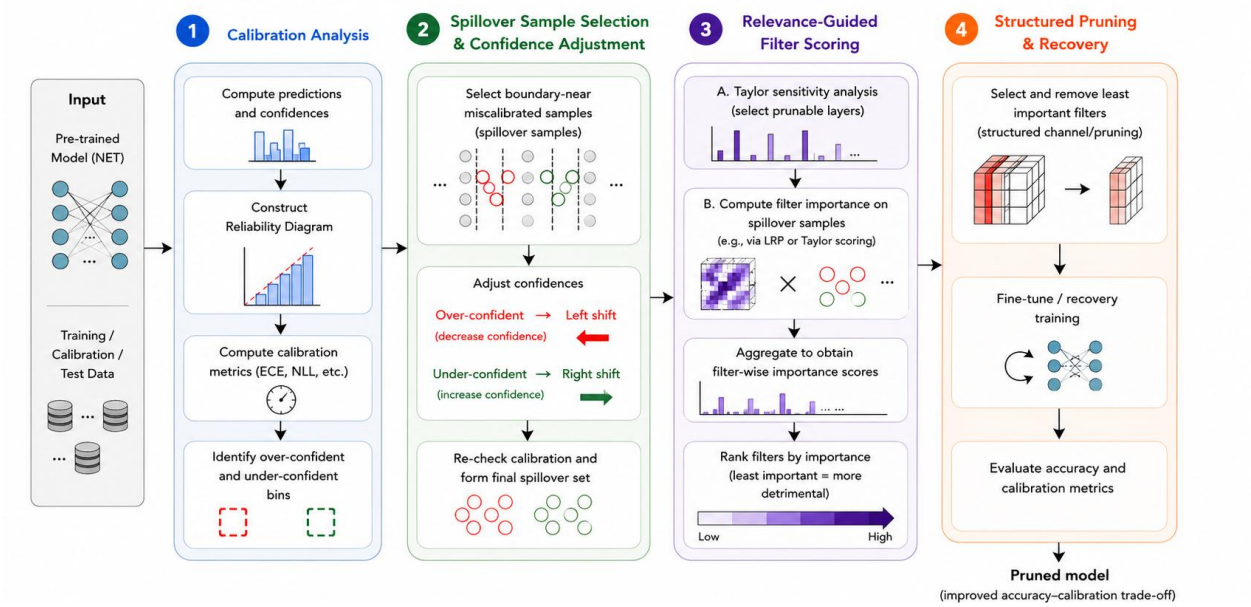

Figure 1: Overview of the proposed calibration-aware sample selection framework. Relevance estimation is performed only on calibration-critical spillover samples, followed by structured filter pruning and recovery fine-tuning.

its hardware-friendly sparsity patterns compared to unstructured or semi-structured pruning (Frankle & Carbin, 2019; Sun et al., 2024; Meng et al., 2020). Structured pruning can be performed before training (Lee et al., 2019; Wang et al., 2020), during training (Liu et al., 2017; Huang & Wang, 2018), or after training; we focus on the post-training (post-hoc) setting, which is especially relevant when compression must be applied to pretrained models without retraining the entire network from scratch (Ma et al., 2023; Yang et al., 2023). Early post-training pruning methods often relied on simple heuristics such as *weight magnitude* (Han et al., 2016b) or *activation statistics* (Li et al., 2017b; He et al., 2018). More recent approaches instead estimate the effect of pruning on loss, e.g., via *loss sensitivity* (Liu et al., 2021; Nonnenmacher et al., 2022), or use *attribution-based criteria* to identify important structures. Separately, entropy-based criteria have been used to mask prunable elements (Liao et al., 2023); unlike our approach, this line of work uses entropy to define the *pruning criterion itself*, rather than to guide which data is used to estimate importance. A related but distinct idea appears in classical active learning, where confidence-based measures guide the selection of *informative samples* for labeling or training (Yang & Loog, 2016); we adopt a comparable sample-selection perspective, but apply it to identify calibration-critical samples for relevance estimation during pruning, rather than for data labeling. Although loss-sensitive methods can provide strong pruning signals, they often require repeated gradient evaluations, layer-wise sensitivity estimation, or additional recovery steps, all of which increase computational cost. In this work, we therefore adopt the well-established LRP attribution method to estimate filter relevance with respect to the model output, following prior attribution-guided pruning approaches in CNNs and ViTs (Yeom et al., 2021; Yu & Xiang, 2023; Voita et al., 2019; Ghadiri et al., 2024).

Uncertainty Calibration. A standard approach to evaluating UC relies on binning techniques. In this setting, predictions are grouped into bins according to their confidence values, and for each bin, the empirical accuracy is compared to the average predicted confidence (Nixon et al., 2019). Plotting per-bin accuracies against their corresponding average confidences yields the reliability diagram (Niculescu-Mizil & Caruana, 2005). Various *binning strategies* have been proposed to improve calibration fidelity, especially under skewed confidence distributions (Nixon et al., 2019). We adopt standard equal-width binning, which partitions the confidence range into fixed-width intervals; while adaptive or equal-frequency binning schemes can improve calibration fidelity under skewed confidence distributions, equal-width bins provide a simpler and more interpretable basis for identifying the calibration-bin boundaries used in our spillover-sample selection. The

idea of directly optimizing the reliability diagram of trained DNNs has led to post-hoc calibration techniques that adjust model outputs at the final layer (Platt, 1999; Zadrozny & Elkan, 2001; 2002; Naeini et al., 2015; Guo et al., 2017), including more recent parameter-free variants (Dimitriadis et al., 2021). Connections between pruning and UC were first drawn by Sun et al. (Sun et al., 2021), who found that non-aggressive unstructured pruning had no adverse effect on UC. Recently, Mitra et al. (Mitra et al., 2024) confirmed similar trends specifically in structured pruning of CNNs. However, only aggressive sparsification degrades UC, and more severely in ViTs than in CNNs (Tao et al., 2023).

Inspired by these findings, our work shifts the focus from measuring UC to leveraging it for structured pruning through calibration-aware sample selection. Although post-hoc calibration methods are computationally efficient, their ability to improve calibration is inherently limited by operating only on the model outputs. More substantial calibration improvements can be achieved using approaches rooted in variational inference or Bayesian approximations, such as Monte Carlo dropout (Kendall & Gal, 2017), evidential learning (Sensoy et al., 2018), deep ensembles (Wilson & Izmailov, 2020), or learnable uncertainty adaptation methods (Kristiadi et al., 2021); however, these methods typically require additional architectural modifications or computational overhead. In contrast, our approach does not optimize an explicit calibration objective. Instead, it exploits reliability diagrams to identify calibration-critical samples, which are subsequently used for relevance estimation and structured filter pruning.

DNN Attribution. At its core, our method relies on attributing an output score – in our case, the predicted class's confidence – to internal units of the DNN. This is the aim of white-box attribution methods from explainable artificial intelligence (Schwalbe & Finzel, 2023). In contrast to costly perturbation-based black-box approaches (Du et al., 2019), these methods rely on DNN-internal gradient, activation, and/or weight information (Zhou et al., 2016; Baehrens et al., 2010; Sundararajan et al., 2017). Amongst them, LRP (Bach et al., 2015; Montavon et al., 2019) is a principled backpropagation-based approach that requires only a single backward pass, and has been successfully employed for pruning in both CNNs and ViTs (Yeom et al., 2021; Yu & Xiang, 2023; Voita et al., 2019; Ghadiri et al., 2024), making it the preferred choice for our method. In summary, although calibration and pruning have each been widely studied in isolation, their combination has received limited attention. Our work offers a first step toward addressing this gap.

3 Background

3.1 Measuring Uncertainty Calibration

UC requires that a model's predicted confidence correspond to the true likelihood that its prediction is correct.

Classification and Confidence. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathcal{X},\ y_i \in \{1, \ldots, K\}$, denote a dataset of N labeled samples, where x_i is the i-th input and y_i is its corresponding ground-truth class label among K classes. Let $f_\theta : \mathcal{X} \to \Delta^{K-1}$ be a multiclass classifier parameterized by θ, where Δ^{K-1} denotes the $(K-1)$-dimensional probability simplex. For a given input x_i, the model outputs a vector of class probabilities $f_\theta(x_i) = \big(f_\theta(x_i)_1, \ldots, f_\theta(x_i)_K\big)$, where $f_\theta(x_i)_k = \mathbb{P}_\theta(Y = k \mid X = x_i)$ and $\sum_{k=1}^{K} f_\theta(x_i)_k = 1$. The predicted class label for x_i is defined as $\hat{y}_i = \arg\max_{k \in \{1,\ldots,K\}} f_\theta(x_i)_k$, and the corresponding prediction confidence is given by $\hat{p}_i = \max_{k \in \{1,\ldots,K\}} f_\theta(x_i)_k$.

Uncertainty Calibration. A classifier is said to be *well calibrated* if its predicted confidence reflects the true probability of correctness. Formally, UC requires $\mathbb{P}(\hat{y}_i = y_i \mid \hat{p}_i = p) = p \quad \text{for all } p \in [0, 1]$.

Reliability Diagrams. A central tool for assessing calibration is the *reliability diagram* (Degroot & Fienberg, 1983; Vaicenavicius et al., 2019), which compares confidence and empirical accuracy between confidence bins. Given a pre-trained DNN f_θ and a test dataset $\mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{N_{\text{test}}}$, predictions are grouped into M bins according to their confidence $\hat{p}_i$, $i = 1, \ldots, N_{\text{test}}$. For each bin B_m, one computes the average confidence $\text{conf}(B_m)$ and accuracy $\text{acc}(B_m)$, and plots the points $\big(\text{conf}(B_m), \text{acc}(B_m)\big)_{m=1}^{M}$. Deviations from the diagonal indicate over- or under-confidence. In our method, the reliability diagram is also used to identify informative miscalibrated regions for spillover sample selection.

Expected Calibration Error (ECE). We summarize calibration using *Expected Calibration Error (ECE)* (Guo et al., 2017),

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N_{\text{test}}} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| .$$

where B_m denotes the set of samples in bin m, $|B_m|$ its cardinality, N_{test} the total number of test samples, and $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ denote the average accuracy and confidence within bin m, respectively.

Negative Log-Likelihood. Alongside ECE, we additionally report the *Negative Log-Likelihood (NLL)*, a proper scoring rule that evaluates the full predictive distribution and strongly penalizes overconfident errors (Minderer et al., 2021):

$$\mathrm{NLL} = -\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \log \hat{p}_i .$$

Together, reliability diagrams, ECE, and NLL provide complementary views of predictive reliability before and after pruning.

3.2 Post-Hoc Recalibration

To disentangle improvements due to pruning from improvements obtainable by calibration alone, we additionally evaluate standard *post-hoc recalibration* methods on the pruned models. Post-hoc recalibration adjusts the predicted confidence scores of an already trained classifier without modifying its underlying feature extractor or decision boundaries.

Temperature Scaling (TS). Temperature scaling (Guo et al., 2017) is a simple and widely used post-hoc calibration method that rescales the logits $z_k(x) = \mathrm{logit}(f_\theta(x)_k)$ of a classifier by a single positive scalar temperature T:

$$f_\theta(x)_k = \frac{\exp(z_k(x)/T)}{\sum_{j=1}^{K} \exp(z_j(x)/T)} .$$

The temperature T is typically optimized on a held-out validation set by minimizing the NLL. Temperature scaling preserves the predicted class labels while adjusting confidence sharpness, making it a strong and lightweight calibration baseline.

Dirichlet Calibration (DC). Dirichlet calibration (Kull et al., 2019) is a more expressive multiclass post-hoc method that learns a parametric transformation of the predicted probability vector $f_\theta(x)$. In its common form, DC applies an affine transformation in log-probability space followed by a softmax normalization:

$$\mathbf{q} = \mathrm{softmax}(W \ \log(f_\theta(x)) + b) .$$

3.3 Relevance Propagation and Feature Importance for Pruning

Our pruning framework leverages LRP (Montavon et al., 2019) to identify filters that contribute minimally to well-calibrated predictions. LRP operates in two stages: (1) a forward pass, during which activations are recorded during standard inference, and (2) a backward relevance-propagation pass, during which the output score is redistributed layer by layer using LRP rules to quantify each unit's contribution. The resulting filter-level relevance scores serve as the pruning criterion in the second stage of our two-stage pipeline (Section **??**): filters with the lowest aggregated relevance are removed, and the pruned model is subsequently fine-tuned to recover performance.

At the core of LRP is the *conservation principle*, which ensures that relevance is preserved across layers. For a neuron i at layer l, its relevance $R_i^{(l)}$ is computed from the relevance $R_j^{(l+1)}$ of neurons j in the next layer:

$$R_i^{(l)} = \textstyle\sum_j \frac{s_{ij}}{\sum_k s_{kj}} R_j^{(l+1)}, \tag{1}$$

where $s_{ij} = a_i w_{ij}$ denotes the contribution from neuron i to neuron j, with a_i being the activation of neuron i and w_{ij} the corresponding weight.

In practice, the ε-LRP rule stabilizes relevance propagation by adding a small stabilizer to the denominator (Bach et al., 2015; Montavon et al., 2019):

$$R_i^{(l)} = \sum_j \frac{s_{ij}}{\sum_k s_{kj} + \epsilon_{\text{LRP}} \operatorname{sign}\left(\sum_k s_{kj}\right)} R_j^{(l+1)}. \tag{2}$$

Here, ϵ_{LRP} reduces numerical instability when the denominator is close to zero. We use ϵ_{LRP} for the LRP stabilizer to distinguish it from the spillover boundary parameter ε.

The γ-LRP rule modifies the local contribution term by emphasizing positive weights,

$$s_{ij}^{\gamma} = a_i\left(w_{ij} + \gamma w_{ij}^{+}\right), \qquad w_{ij}^{+} = \max(w_{ij}, 0),$$

before applying the same relevance redistribution principle (Montavon et al., 2019). Composite LRP combines different propagation rules across layer groups and has been commonly used in CNN attribution analyses; we adopt it for VGG-16, following prior LRP-based pruning studies (Montavon et al., 2019; Yeom et al., 2021). For the remaining architectures (ResNet-34/56, DenseNet-121, and DeiT), we apply the ε-LRP rule uniformly across layers.

For filter pruning, relevance is propagated from the predicted-class logit and aggregated over samples at the filter level. Filters with the lowest aggregated relevance are interpreted as contributing least to the selected output behavior and are therefore selected as pruning candidates.

4 Calibration-Aware Sample Selection for Pruning

Algorithm 1: Calibration-Aware Relevance-Guided Filter Pruning

Input: Pretrained model f_θ, datasets $D_{\text{train}}, D_{\text{val}}, D_{\text{test}}$, pruning ratio r, number of bins B, search space Ω over $(\delta_{\text{over}}, \delta_{\text{under}}, \varepsilon)$

Output: Selected pruned model $f_{\theta^\star}$ and hyperparameters $(\delta^\star_{\text{over}}, \delta^\star_{\text{under}}, \varepsilon^\star)$

Initialize optimizer over Ω;
while *budget not exhausted* **do**
  Sample $(\delta_{\text{over}}, \delta_{\text{under}}, \varepsilon) \in \Omega$;
  `// Step 1: Calibration-aware spillover sample selection`
  $(\hat{y}, c, b) \leftarrow \text{Bin}(f_\theta, D_{\text{test}}, B)$;
  $\mathcal{B}_{\text{over}}, \mathcal{B}_{\text{under}} \leftarrow \text{MisBins}(\hat{y}, y, c, b)$;
  $D_{\text{spill}} \leftarrow \text{Spill}(D_{\text{test}}, \mathcal{B}_{\text{over}}, \mathcal{B}_{\text{under}}, \varepsilon)$;
  $c' \leftarrow \text{Adjust}(c, D_{\text{spill}}, \delta_{\text{over}}, \delta_{\text{under}}, \varepsilon)$; `// diagnostic only: verifies` D_{spill} `is miscalibration-responsible; not used in Steps 2-5`
  `// Step 2: Taylor-based layer screening`
  for *layer* ℓ **do**
    $S_\ell = \mathbb{E}_{(x,y)\sim D_{\text{spill}}}\left[\left\|\frac{\partial \mathcal{L}(f_\theta(x), y)}{\partial W_\ell} \odot W_\ell\right\|_1\right]$;
  $\mathcal{L}_{\text{prune}} = \text{KneePoint}(\{S_\ell\}_\ell)$;
  `// Step 3: LRP-based filter relevance`
  for $\ell \in \mathcal{L}_{\text{prune}}$ **do**
    for *filter* k *in layer* ℓ **do**
      $R_{\ell,k} = \frac{1}{|D_{\text{spill}}|}\sum_{(x_i,y_i)\in D_{\text{spill}}} \left|\text{LRP}_{\ell,k}(f_\theta, x_i)\right|$;
  `// Step 4: Structured pruning and fine-tuning`
  $\mathcal{P}_\ell = \text{Bottom}_{r\%}(\{R_{\ell,k}\}_k)$ for each $\ell \in \mathcal{L}_{\text{prune}}$;
  $f_{\theta'} = \text{Prune}(f_\theta, \{\mathcal{P}_\ell\}_{\ell\in\mathcal{L}_{\text{prune}}})$;
  $f_{\theta'} = \text{FineTune}(f_{\theta'}, D_{\text{train}})$;
  `// Step 5: Validation`
  $J = \text{ValObjective}(f_{\theta'}, D_{\text{val}})$;
  Update optimizer with J;
Select $(\delta^\star_{\text{over}}, \delta^\star_{\text{under}}, \varepsilon^\star)$ and $f_{\theta^\star}$ with best validation objective;
Evaluate $f_{\theta^\star}$ on D_{test};

We propose a pruning framework that integrates UC diagnostics with attribution-based relevance scores to guide structured filter pruning, as outlined in Algorithm 1. Starting from a pretrained model, we first identify informative miscalibrated samples via reliability analysis to form a spillover set, and separately verify via a diagnostic confidence adjustment that these samples are indeed responsible for the observed miscalibration (Sec. 20). We then perform Taylor-based sensitivity analysis to identify prunable layers (Sec. 20), compute LRP-based filter relevance scores on the selected spillover samples (Sec. 3.3), and use the aggregated relevance to drive structured filter pruning (Sec. 20). The resulting pruned model is subsequently fine-tuned and validated to recover predictive performance while preserving calibration.

Step 1: Calibration-Aware Spillover Sample Selection and Confidence Adjustment. We begin by analyzing the calibration behavior of a pretrained model f_θ on the test set $D_{\text{test}} = \{(x_i, y_i)\}_{i=1}^N$. For each sample, we compute the predicted confidence $c_i = \max_k f_\theta(x_i)_k$ and predicted label $\hat{y}_i = \arg\max_k f_\theta(x_i)_k$. The samples are partitioned into B equal-width bins over the interval $[0, 1]$ based on their confidence values, where bin b is defined as

$$I_b = \left(\frac{b-1}{B}, \frac{b}{B}\right], \quad b = 1, \ldots, B.$$

For each bin b, we compute the average confidence and empirical accuracy:

$$\text{conf}(b) = \frac{1}{|I_b|} \sum_{i \in I_b} c_i, \quad \text{acc}(b) = \frac{1}{|I_b|} \sum_{i \in I_b} \mathbf{1}(\hat{y}_i = y_i).$$

Bins are categorized as *overconfident* if $\text{conf}(b) > \text{acc}(b)$, and *underconfident* if $\text{conf}(b) < \text{acc}(b)$. From both bin types, we consider wrongly predicted samples (x_i, y_i) with $\hat{y}_i \neq y_i$, which form the basis for the spillover set. These samples are selected based on their proximity to bin boundaries using a selection procedure (cf. Algorithm 1), yielding the subset

$$D_{\text{spill}} = D_{\text{over}}^{\text{wrong}} \cup D_{\text{under}}^{\text{wrong}}.$$

To diagnostically assess whether these samples are indeed responsible for the observed miscalibration, we apply an edge-aware adjustment to their confidence values, without retraining or otherwise modifying the model. Let $x_i \in I_b = \left(\frac{b-1}{B}, \frac{b}{B}\right]$, with lower and upper bin boundaries $l_b = \frac{b-1}{B}$ and $u_b = \frac{b}{B}$. The tolerance ε defines how close a sample must lie to a bin boundary to be treated as a boundary case during adjustment.

For spillover samples originating from overconfident bins, the adjusted confidence is defined as

$$c_i' = \begin{cases} c_i - \delta_{\text{over}}, & \text{if } |c_i - u_b| \leq \varepsilon, \\ l_b - \delta_{\text{over}}, & \text{otherwise,} \end{cases}$$

i.e., samples close to the upper bin boundary are treated as already representative of the bin's overconfidence and are only mildly perturbed within the bin, while all other samples are explicitly pushed across the lower boundary into the adjacent, less-confident bin, so as to maximally test whether removing their (miscalibrated) confidence mass from this bin improves calibration.

For spillover samples from underconfident bins, the adjustment is given by

$$c_i' = \begin{cases} c_i + \delta_{\text{under}}, & \text{if } |c_i - l_b| \leq \varepsilon, \\ u_b + \delta_{\text{under}}, & \text{otherwise,} \end{cases}$$

i.e., samples close to the lower boundary are perturbed within the bin, while others are shifted across the upper boundary into the adjacent, more-confident bin, following the same rationale.

This edge-aware adjustment mechanism ensures that confidence updates either preserve bin membership (for boundary-near samples already representative of the bin) or explicitly induce spillover into neighboring bins, thereby enabling controlled manipulation of bin-wise calibration statistics for diagnostic purposes.

The hyperparameters ε, δ_{over}, and δ_{under} control the sensitivity to bin boundaries and the magnitude of confidence adjustments, and are specified in the experimental setup. This confidence-adjustment step serves solely as a diagnostic check confirming that the selected spillover samples are responsible for the observed miscalibration; the *unadjusted* spillover set D_{spill} is what is used as the calibration-informed input for the downstream pruning stages (Steps 2–4).

Step 2: Taylor Sensitivity Analysis: Layer Screening. Given the calibration-informed spillover set D_{spill}, we identify prunable layers via a first-order Taylor sensitivity analysis **?**. Let W_ℓ denote the learnable parameters of layer ℓ. We estimate the expected magnitude of the first-order loss change with respect to the layer parameters W_ℓ as

$$S_\ell = \mathbb{E}_{(x_i,y_i)\sim D_{\text{spill}}}\left[\left\|\frac{\partial \mathcal{L}(f_\theta(x_i), y_i)}{\partial W_\ell} \odot W_\ell\right\|_1\right],$$

Layers are ranked in descending order of S_ℓ, where larger scores indicate higher sensitivity to pruning. We apply a knee-point heuristic to the sorted scores to obtain a threshold τ_{knee}. Layers above this threshold are treated as sensitive and excluded from pruning,

$$\mathcal{L}_{\text{sens}} = \{\ell : S_\ell \geq \tau_{\text{knee}}\},$$

while the remaining layers are considered prunable,

$$\mathcal{L}_{\text{prune}} = \{\ell : S_\ell < \tau_{\text{knee}}\}.$$

Subsequent LRP-based filter relevance analysis is restricted to layers in $\mathcal{L}_{\text{prune}}$.

Step 3: LRP-Based Filter Relevance on Spillover Samples. For the prunable layers $\mathcal{L}_{\text{prune}}$, we compute relevance scores for the corresponding prunable structural units using LRP on the calibration-informed spillover set D_{spill}. For CNNs (VGG-16, ResNet-34/56, DenseNet-121), the prunable unit is the convolutional filter; for DeiT, the prunable unit is an output neuron of the MLP (feed-forward) linear layers within each transformer block. For each sample $(x_i, y_i) \in D_{\text{spill}}$, relevance is propagated from the model output (logits) back to the target layer. For convolutional filters, we spatially average the absolute relevance values per filter; for DeiT MLP units, relevance is taken directly at the corresponding output neuron. In both cases, scores are normalized within the layer and aggregated across spillover samples to obtain a mean relevance score per unit. Units with the lowest aggregated relevance are interpreted as contributing least to the model's behavior on these calibration-critical mispredictions and therefore form the basis for the structured pruning decisions.

Step 4: Structured Pruning. Using the aggregated filter relevance scores, we construct a structured pruned model by removing the bottom $r\%$ of filters in each prunable layer $\ell \in \mathcal{L}_{\text{prune}}$. The resulting compact model is then fine-tuned in the subsequent stage to recover predictive performance.

Step 5–6: Fine-Tuning, Validation, and Hyperparameter Selection. Following structured pruning, each pruned model is fine-tuned on the training set D_{train} to recover predictive performance; for DeiT, the pruning mask is re-applied after each optimizer step to preserve the pruned structure throughout fine-tuning. For each pruning ratio r, all candidate hyperparameter configurations $(\delta_{\text{over}}, \delta_{\text{under}}, \varepsilon)$ are evaluated on the validation set D_{val} using accuracy, ECE, and NLL. Let $\mathcal{C}_r$ denote the set of candidate models at pruning ratio r. We compute the Pareto-optimal set $\mathcal{P}_r \subseteq \mathcal{C}_r$ with respect to the objectives $(\text{Acc}(f), -\text{ECE}(f), -\text{NLL}(f))$, i.e., the set of candidates in $\mathcal{C}_r$ not dominated by any other candidate on all three objectives simultaneously. From this set, we select the candidate with the highest validation accuracy,

$$f_r^\star = \arg\max_{f\in\mathcal{P}_r} \text{Acc}(f),$$

with ties broken by lower ECE and then lower NLL. The selected hyperparameter configuration and corresponding pruned model are retained for each pruning ratio and finally evaluated on D_{test}.

5 Experimental Setup

Dataset & Models. We evaluate our method on three standard image classification benchmarks: **CIFAR-10**, **CIFAR-100** (Krizhevsky & Hinton, 2009), and **ImageNet** (Deng et al., 2009). To ensure broad applicability, we include both convolutional and transformer-based architectures, covering a range of model sizes and complexities. Following prior pruning benchmarks Li et al. (2017a); Fang et al. (2023), we evaluate on the following model–dataset combinations.

On CIFAR-10 and CIFAR-100, we use **VGG-16** (Li et al., 2017a), **ResNet-56** Li et al. (2017a), and **DenseNet-121** (Fang et al., 2023). On CIFAR-100 additionally, we evaluate **DeiT-Tiny** (Touvron et al., 2021). On ImageNet, we evaluate **ResNet-34** (Li et al., 2017a).

Training Setup. All models are initialized from ImageNet-pretrained weights from the PyTorch model zoo (Paszke et al., 2019). For CIFAR-10 and CIFAR-100, models are retrained for 100 epochs using SGD with momentum 0.9, weight decay 10^{-4}, and initial learning rate 0.001, following standard CIFAR training schedules (He et al., 2016; Huang et al., 2017). For ImageNet, we use the corresponding ImageNet-pretrained models without additional dataset adaptation before pruning.

LRP Configurations. For VGG-16, we follow the composite LRP configuration used in prior LRP-based CNN attribution and pruning studies, combining γ-LRP and ε-LRP rules (Montavon et al., 2019; Yeom et al., 2021). For ResNet-56, ResNet-34, DenseNet-121, and DeiT-Tiny, we use ε-LRP throughout the network, following its standard use as a stabilizing propagation rule and to maintain a uniform configuration across residual, densely connected, and transformer-based architectures (Bach et al., 2015; Montavon et al., 2019; Achtibat et al., 2024).

The candidate components for relevance scoring are defined architecture-wise, following common structured-pruning practice for the corresponding model families (Li et al., 2017b; Molchanov et al., 2019; **?**). For VGG-16, we consider convolutional layers as candidate components. For ResNet-34 and ResNet-56, we consider the first convolutional layer, `conv1`, of each residual block. For DenseNet-121, we consider `conv1` and `conv2` in each bottleneck unit. For DeiT-Tiny, we consider the MLP layers, `fc1` and `fc2`, within each transformer block. After Taylor-based layer screening, LRP-based relevance is computed only for the selected prunable layers.

Evaluation Metrics. We evaluate each model using metrics that capture predictive performance, uncertainty estimation, and calibration quality. Predictive performance is measured by top-1 classification accuracy on the test set. Uncertainty estimation is assessed using NLL, while calibration quality is evaluated using ECE. For ECE, we use equal-width binning with 10 uniformly spaced bins over the confidence interval $[0, 1]$. In addition to the raw model outputs, we also report post-hoc recalibrated metrics using TS and DC. Both post-hoc calibrators are fitted on the validation set and evaluated on the test set. We report the corresponding recalibrated NLL and ECE values, denoted as TS-NLL, TS-ECE, DC-NLL, and DC-ECE.

Baseline Pruning Techniques. We compare our calibration-aware pruning method against two conceptually distinct baselines: a magnitude-based filter pruning method and an explainability-driven pruning approach. **L1-Norm Pruning** (Li et al., 2017b) removes filters with the smallest weight magnitudes, operating under the assumption that lower-magnitude filters contribute less to the output. **LRP Pruning** (Yeom et al., 2021) leverages LRP to assign relevance scores to filters based on their contribution to the final prediction, pruning those with the lowest relevance to reduce model complexity.

Hyperparameter Optimization. We search $\delta_{\text{over}}, \delta_{\text{under}} \in [0.005, 0.050]$ and $\varepsilon \in [0.005, 0.010]$. Here, δ_{over} and δ_{under} control the absolute confidence-shift magnitude for overconfident and underconfident spillover samples, respectively, while ε defines the absolute bin-boundary tolerance for identifying boundary-near samples. With $B = 10$ equal-width bins, the bin width is $1/B = 0.1$; all three hyperparameters are specified in absolute confidence units on $[0, 1]$, rather than as fractions of the bin width. For computational tractability and fair comparison, we keep the hyperparameter search space fixed across all evaluated architectures within each dataset.

6 Results

We evaluate our calibration-aware relevance-guided pruning method across three datasets, comparing it with existing approaches using both predictive and calibration metrics. Results are grouped by dataset for clarity. Each graph shows the effect of increasing pruning ratios on accuracy, ECE, NLL, and re-calibrated ECE and NLL. We use the same binning and evaluation procedure across all methods to ensure fair comparison.

Table 1: Structured pruning performance on CIFAR-10 across representative sparsity levels. Accuracy (%) is higher-is-better (↑), while ECE and NLL are lower-is-better (↓).

Model	Method	10%			20%			40%			60%			80%			90%		
		Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL
VGG-16	L1 Norm	**93.72**	4.42	0.34	**93.27**	4.66	0.35	90.61	4.24	0.32	**89.84**	5.31	0.39	84.51	6.09	0.51	80.62	4.55	0.58
	LRP Pruning	92.70	4.60	0.32	92.15	4.88	0.34	90.65	4.05	0.33	86.65	**3.21**	0.39	81.50	4.02	0.54	10.00†	**0.31**†	2.30
	Ours	93.28	**2.11**	**0.22**	93.16	**2.04**	**0.21**	**92.21**	**2.70**	**0.25**	89.71	3.80	**0.33**	**85.37**	**2.99**	**0.45**	**82.75**	2.83	**0.51**
ResNet-56	L1 Norm	93.13	3.87	0.29	**92.50**	4.55	0.31	**92.31**	4.54	0.32	**90.67**	4.46	**0.32**	**85.92**	3.85	**0.42**	80.12	2.86	**0.51**
	LRP Pruning	**93.23**	4.08	0.28	**92.50**	4.45	**0.30**	90.95	5.24	0.37	89.43	6.22	0.44	84.46	6.44	0.53	**80.98**	5.89	0.61
	Ours	92.27	**3.55**	**0.27**	91.91	**4.41**	0.31	91.33	**4.19**	**0.31**	89.83	**2.78**	0.48	85.57	**3.36**	0.44	80.21	**1.78**	0.56
DenseNet-121	L1 Norm	95.48	2.57	0.18	95.16	2.69	0.19	94.73	2.94	0.20	**94.54**	2.88	0.21	93.50	2.79	0.22	**92.22**	2.53	**0.24**
	LRP Pruning	**95.55**	**2.09**	**0.17**	95.20	2.28	0.19	94.81	2.83	0.20	**94.54**	**2.10**	0.19	**93.55**	2.11	**0.20**	91.68	**1.77**	0.25
	Ours	95.15	2.66	0.20	**95.46**	**2.27**	**0.18**	**95.11**	**2.72**	0.20	94.48	2.25	**0.18**	92.86	**1.96**	0.22	91.25	2.16	0.27

†LRP Pruning at 90% sparsity on VGG-16 collapses to near-random accuracy; its low ECE reflects degenerate, uniformly low-confidence predictions rather than genuine calibration, not a meaningful calibration result.

CIFAR-10. Tables 1 and 2 summarize the pruning and post-hoc calibration results on CIFAR-10. Across VGG-16 and ResNet-56, the proposed calibration-aware pruning consistently achieves the lowest calibration error while maintaining competitive, and often superior, predictive accuracy. For VGG-16, our method improves both calibration and accuracy at moderate and high pruning levels, reducing ECE by up to 56% relative to L1-norm and LRP pruning (e.g., at 20% sparsity), while achieving the highest accuracy at 40%, 80%, and 90% sparsity; LRP is briefly ahead on ECE only at 60% sparsity. Similar trends are observed for ResNet-56, where our approach consistently yields the lowest ECE across all pruning levels without substantial loss in accuracy, trailing L1-Norm only slightly at higher sparsity. For DenseNet-121, results are more mixed: our method exceeds both baselines in accuracy at moderate sparsity (20%, 40%) but trails L1-Norm at low and high sparsity levels, and LRP pruning achieves marginally better calibration at 10%, 60%, and 80% sparsity, while our method achieves the lowest ECE at the remaining levels (20%, 40%, 90%) – suggesting that DenseNet-121's dense connectivity already provides relatively stable feature representations that leave less headroom for calibration-aware sample selection to improve upon.

Post-hoc recalibration further reduces calibration error for most methods, particularly under Temperature Scaling. However, its effect on relative ranking is architecture-dependent: for VGG-16, our method retains the lowest ECE and NLL under Raw, TS, and DC alike, indicating that the calibration-aware sample selection improves the model's intrinsic calibration rather than merely making it easier to recalibrate. For ResNet-56 and DenseNet-121, Dirichlet Calibration occasionally overturns this ranking in favor of the baselines – for instance, L1-Norm achieves the lowest DC-ECE for ResNet-56 at every tested sparsity level. This is consistent with DC's greater expressiveness relative to temperature scaling: because DC learns a full affine transformation rather than a single scalar, it is more prone to overfitting the calibration split, particularly at higher sparsity where pruning has already altered the model's logit distribution. We also note that DC directly optimizes a likelihood-based objective rather than ECE, so its improvements are not guaranteed to translate into lower ECE in every setting. Given this variability, we do not claim that Dirichlet Calibration is uniformly beneficial; we report it as an additional post-hoc baseline and observe that Temperature Scaling is the more stable recalibration method across our experiments.

CIFAR-100. Table 3 shows that the proposed calibration-aware pruning strategy maintains competitive predictive performance while improving UC across diverse architectures. On ResNet-56, our method achieves the highest accuracy at low-to-moderate pruning levels (10–40%), while LRP pruning generally yields lower ECE at higher sparsity levels (60% and 90%), with L1-norm pruning ahead only at 80%. For DenseNet-121, our approach preserves the highest accuracy from 40% pruning onward, while LRP pruning consistently achieves the lowest calibration error. On DeiT-Tiny, our method consistently achieves the highest accuracy across all pruning levels and improves raw calibration at low-to-moderate sparsity, reducing ECE from 11.30% to 10.43% and NLL from 0.91 to 0.85 at 10% pruning compared with L1 pruning. At higher pruning levels (80–90%), however, this trend reverses: our method's calibration deteriorates substantially more than the baselines', with ECE rising to 13.68% and 11.56% at 80% and 90% respectively – roughly double and triple

Table 2: Effect of post-hoc recalibration under structured pruning on CIFAR-10. Each cell reports ECE (%) and NLL as Raw / TS / DC at representative pruning levels. Lower is better (↓). Best TS and DC values at each pruning level are highlighted in bold.

Model	Method	10%		50%		90%	
		ECE (Raw/TS/DC)↓	NLL (Raw/TS/DC)↓	ECE (Raw/TS/DC)↓	NLL (Raw/TS/DC)↓	ECE (Raw/TS/DC)↓	NLL (Raw/TS/DC)↓
VGG-16	L1	4.42 / 3.55 / 1.12	0.34 / 0.26 / **0.19**	4.78 / 3.54 / 0.97	0.33 / 0.28 / 0.25	4.55 / 3.30 / 1.61	0.58 / 0.57 / 0.53
	LRP	4.60 / 4.98 / 4.22	0.32 / 0.36 / 0.30	3.95 / 5.00 / 5.60	0.39 / 0.41 / 0.43	0.31† / 0.31† / 0.52†	2.30 / 2.30 / 2.30
	Ours	**2.11** / **1.65** / **0.83**	**0.22** / **0.21** / 0.20	**3.10** / **2.20** / **0.76**	**0.26** / **0.24** / **0.24**	**2.83** / **1.99** / **1.13**	**0.51** / **0.50** / **0.49**
ResNet-56	L1	3.87 / 3.98 / **2.98**	0.29 / 0.29 / **0.25**	4.46 / 3.68 / **2.49**	**0.30** / 0.30 / **0.28**	2.86 / 1.41 / **0.98**	**0.51** / **0.50** / **0.50**
	LRP	4.10 / 4.11 / 4.46	0.28 / 0.28 / 0.32	5.59 / 6.60 / 6.61	0.40 / 0.49 / 0.50	5.92 / 7.08 / 8.43	0.61 / 0.63 / 0.63
	Ours	**3.55** / **2.80** / 4.37	**0.27** / **0.25** / 0.32	**4.21** / **3.35** / 4.93	0.32 / **0.29** / 0.37	**1.78** / **1.15** / 1.24	0.86 / 0.86 / 0.82
DenseNet-121	L1	2.57 / 0.58 / **0.57**	0.18 / **0.15** / **0.15**	2.73 / **0.32** / 0.81	0.20 / 0.17 / **0.17**	2.53 / 0.83 / 1.15	**0.24** / **0.23** / **0.23**
	LRP	**2.09** / 2.99 / 2.50	**0.17** / 0.23 / 0.19	2.30 / 3.03 / 2.92	0.19 / 0.23 / 0.22	**1.77** / 2.71 / 2.64	0.25 / 0.26 / 0.26
	Ours	2.66 / **0.32** / 2.87	0.20 / 0.23 / 0.21	**2.13** / 0.52 / **0.71**	**0.18** / **0.16** / **0.17**	2.16 / **0.41** / **0.65**	0.27 / 0.26 / 0.27

†Model collapse at 90% pruning; see Table 1.

Table 3: Structured pruning performance on CIFAR-100 across representative sparsity levels. Conventions follow Table 1.

Model	Method	10%			20%			40%			60%			80%			90%		
		Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL
VGG-16	L1 Norm	73.95	**7.24**	**0.95**	71.43	7.14	1.04	64.54	**3.40**	1.24	58.25	3.29	1.50	42.06	1.96	2.14	19.20	2.25	3.32
	LRP Pruning	**74.16**	13.04	1.18	**73.58**	12.35	1.12	**71.07**	12.06	1.18	**66.84**	8.77	**1.21**	**44.42**	1.89	**2.08**	**20.17**	6.01	**3.32**
	Ours	73.24	7.48	0.96	72.50	**6.41**	**0.98**	68.52	5.55	**1.12**	65.36	**1.68**	1.58	30.76	**1.82**	2.67	13.00	**1.82**	3.62
ResNet-56	L1 Norm	68.58	**5.87**	1.15	67.19	**4.71**	1.19	58.97	**3.48**	1.45	46.68	**2.72**	1.97	39.69	**2.03**	2.27	34.59	2.21	2.47
	LRP Pruning	70.86	8.37	**1.09**	69.49	9.05	**1.15**	67.23	9.10	1.25	**63.38**	8.07	**1.34**	**57.80**	**4.11**	**1.51**	**51.69**	1.22	**1.77**
	Ours	**71.61**	11.10	1.15	**70.45**	11.47	1.21	**67.47**	8.50	**1.23**	63.25	**8.01**	1.39	50.61	5.27	1.88	31.52	**0.82**	2.61
DenseNet-121	L1 Norm	77.86	11.18	0.97	77.38	10.96	0.96	76.60	9.77	0.93	75.94	9.54	0.96	73.87	7.21	0.95	71.56	6.10	1.02
	LRP Pruning	**78.44**	**9.54**	**0.90**	**78.46**	**8.53**	**0.88**	77.40	**9.09**	**0.90**	76.98	**8.70**	**0.90**	**74.69**	**6.67**	**0.92**	71.56	**5.06**	**1.01**
	Ours	78.05	11.52	0.99	78.08	11.17	0.99	**77.59**	11.33	0.97	**77.20**	8.97	**0.90**	74.68	7.86	0.93	**71.91**	7.36	**1.01**
DeiT-Tiny	L1 Norm	83.65	11.30	0.91	82.23	12.05	0.94	79.61	12.75	1.02	75.17	**10.36**	**1.00**	68.94	**6.48**	**1.16**	63.53	**4.08**	**1.31**
	LRP Pruning	83.91	11.34	0.92	82.76	12.00	0.93	80.85	12.25	0.97	76.05	12.02	1.04	67.64	7.19	1.20	62.35	4.25	1.36
	Ours	**84.74**	**10.43**	**0.85**	**83.98**	**10.89**	**0.88**	**82.23**	**11.58**	**0.95**	**78.01**	13.26	1.09	**70.46**	13.68	1.30	**65.39**	11.56	1.42

the ECE of L1-norm and LRP pruning at the same sparsity levels – even though our method continues to retain the strongest predictive accuracy.

Table 4 further shows that post-hoc recalibration substantially improves calibration for all pruning strategies. Temperature Scaling consistently produces the largest ECE reductions, whereas Dirichlet Calibration often provides competitive or lower NLL. Despite these improvements, the relative trends remain largely unchanged, indicating that the proposed pruning strategy and post-hoc calibration provide complementary benefits.

ImageNet. Unlike the CIFAR experiments, where six representative pruning levels are reported, the ImageNet results are summarized at four representative operating points (10%, 20%, 50%, and 90%), corresponding to low, moderate, and extreme sparsity regimes. This provides a compact summary of the large-scale ImageNet experiments while capturing the key trends across the pruning spectrum. As shown in Table 5, at low pruning levels (10–20%), our method attains the lowest NLL and, at 20%, the lowest ECE as well, while L1-norm pruning retains a marginally higher accuracy at both levels. At the intermediate pruning level (50%), our method continues to achieve the lowest calibration error and NLL, whereas L1 pruning retains a marginally higher accuracy. Under extreme pruning (90%), LRP pruning exhibits the strongest overall performance across accuracy, ECE, and NLL alike, suggesting that relevance-based filter ranking becomes increasingly beneficial when the network capacity is substantially reduced. Table 6 further shows that Temperature Scaling and Dirichlet Calibration generally improve calibration across pruning methods, though not uniformly: at 50% sparsity, both recalibration methods increase ECE for L1-norm pruning, and Dirichlet Calibration similarly increases ECE for our method, consistent with the DC instability observed on CIFAR-10. Nevertheless, our approach maintains the lowest recalibrated NLL at low pruning levels (10%), indicating that calibration-aware sample selection and post-hoc recalibration provide complementary benefits in this regime.

Table 4: Effect of post-hoc recalibration under structured pruning on CIFAR-100. Conventions follow Table 2.

Model	Method	10%		50%		90%	
		ECE (Raw / TS / DC)↓	NLL (Raw / TS / DC)↓	ECE (Raw / TS / DC)↓	NLL (Raw / TS / DC)↓	ECE (Raw / TS / DC)↓	NLL (Raw / TS / DC)↓
VGG-16	L1 Norm	**7.24** / 5.32 / **6.49**	**0.949** / **0.922** / **0.942**	**2.52** / **2.31** / **2.47**	1.337 / 1.338 / 1.301	2.25 / 1.92 / **1.76**	**3.319** / 3.317 / **3.287**
	LRP Pruning	13.04 / 14.78 / 15.59	1.175 / 1.300 / 1.346	10.12 / 12.52 / 13.08	**1.172** / **1.253** / **1.235**	6.01 / 3.39 / 1.97	3.322 / **3.296** / 3.299
	Ours	7.48 / **5.31** / 20.91	0.964 / 0.933 / 1.925	5.04 / 3.27 / 21.14	**1.301** / 1.288 / 2.075	**1.82** / **1.66** / 1.79	3.620 / 3.619 / 3.550
ResNet-56	L1 Norm	**5.87** / **6.68** / **8.98**	1.150 / **1.160** / **1.338**	**2.49** / **1.96** / **6.52**	1.683 / 1.679 / 1.834	2.21 / 1.00 / **4.20**	2.466 / 2.460 / **2.534**
	LRP Pruning	8.37 / 15.03 / 22.14	**1.088** / 1.331 / 2.173	8.54 / 17.46 / 26.51	**1.296** / 1.617 / 2.682	1.22 / 5.37 / 30.82	**1.768** / **1.793** / 3.198
	Ours	11.10 / 12.89 / 15.45	1.154 / 1.225 / 1.403	8.83 / 9.18 / 14.54	**1.287** / **1.294** / **1.497**	**0.43** / **0.66** / **2.63**	2.614 / 2.612 / 2.665
DenseNet-121	L1 Norm	11.64 / 1.82 / 11.64	1.15 / **0.79** / 1.15	11.99 / 1.57 / 11.95	1.20 / 0.85 / 1.20	10.88 / 1.45 / **10.88**	1.30 / 0.99 / 1.30
	LRP Pruning	**9.54** / 10.65 / **11.93**	**0.90** / 0.95 / **1.03**	**8.70** / 12.65 / **10.61**	**0.90** / 1.08 / **1.04**	**5.06** / 6.00 / 11.11	**1.01** / 1.02 / **1.24**
	Ours	11.52 / **1.72** / 14.11	0.99 / **0.79** / 1.31	8.97 / **1.09** / 16.64	**0.90** / **0.80** / 1.47	7.36 / **1.09** / 14.05	**1.01** / **0.96** / 1.42
DeiT-Tiny	L1 Norm	11.30 / 10.95 / **2.04**	0.91 / 0.87 / 0.76	12.89 / 7.74 / **1.58**	**1.02** / 0.83 / **0.87**	**4.08** / 1.63 / **2.52**	**1.31** / 1.31 / **1.55**
	LRP Pruning	11.34 / 4.10 / 2.99	0.92 / 0.67 / 0.76	**12.47** / 3.91 / 1.81	1.03 / 0.83 / 0.96	4.25 / 2.00 / 6.85	1.36 / 1.35 / 1.86
	Ours	**10.43** / **3.09** / 3.17	**0.85** / **0.63** / **0.72**	**12.47** / **3.28** / 2.15	1.03 / **0.78** / 0.91	11.56 / **1.49** / 7.22	1.42 / **1.30** / 1.81

Table 5: Structured pruning performance on ImageNet across representative sparsity levels.

Model	Method	10%			20%			50%			90%		
		Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL	Acc	ECE	NLL
ResNet-34	L1 Norm	73.26	2.44	1.08	72.42	2.26	1.12	**68.78**	1.12	1.26	49.61	4.51	2.18
	LRP Pruning	71.88	**2.19**	1.13	71.04	2.29	1.16	67.59	1.14	1.31	**52.28**	**3.65**	**2.06**
	Ours	**73.09**	2.34	**1.07**	**72.15**	**2.04**	**1.10**	68.29	**1.08**	**1.25**	49.67	5.43	2.18

Table 6: Effect of post-hoc recalibration under structured pruning on ImageNet

Model	Method	10%		50%		90%	
		ECE (Raw / TS / DC)↓	NLL (Raw / TS / DC)↓	ECE (Raw / TS / DC)↓	NLL (Raw / TS / DC)↓	ECE (Raw / TS / DC)↓	NLL (Raw / TS / DC)↓
ResNet-34	L1 Norm	2.44 / 1.56 / 1.96	1.07 / 1.08 / 1.09	**1.12** / 1.15 / 1.16	**1.26** / **1.26** / **1.25**	4.51 / 1.92 / 4.08	2.18 / 2.18 / 2.16
	LRP Pruning	**2.19** / 1.35 / **1.57**	1.13 / 1.13 / 1.11	1.14 / **1.02** / **1.05**	1.31 / 1.31 / 1.29	**3.65** / **1.79** / **3.25**	**2.06** / **2.05** / **2.02**
	Ours	2.34 / **1.27** / 1.69	**1.07** / **1.07** / **1.07**	1.28 / 1.08 / 1.12	1.29 / 1.29 / 1.28	5.43 / 2.30 / 4.96	2.18 / 2.18 / 2.16

7 Conclusion

This work introduced a calibration-aware sample selection strategy for structured filter pruning that explicitly incorporates UC into the pruning process. Rather than estimating filter importance from the entire validation set, the proposed approach focuses relevance estimation on calibration-critical spillover samples identified from reliability diagrams. Combined with LRP-based filter ranking, this enables pruning decisions that account for both predictive performance and confidence reliability.

Extensive experiments on CNNs and a vision transformer for the CIFAR-10, CIFAR-100, and ImageNet datasets demonstrate that calibration-aware sample selection improves UC while maintaining competitive predictive accuracy relative to standard L1-norm and LRP-based structured pruning, particularly at low-to-moderate sparsity levels. Furthermore, the proposed method remains largely complementary to post-hoc recalibration: Temperature Scaling consistently improves calibration further, though Dirichlet Calibration's benefits are less consistent, occasionally increasing calibration error at higher sparsity levels.

These results highlight UC as an effective optimization objective for structured model compression, extending pruning beyond traditional accuracy-driven criteria. More broadly, the proposed framework demonstrates that incorporating reliability information into pruning decisions can improve the trustworthiness of compressed deep neural networks without compromising their deployment efficiency.

Future work will investigate calibration-aware pruning for larger-scale vision foundation models and multi-modal architectures, jointly optimizing calibration, robustness, and computational efficiency under hardware-aware deployment constraints.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11: 1803–1831, August 2010. ISSN 1532-4435.

Simon Burton. The Path to Safe Machine Learning for Automotive Applications. Technical report, SAE International, October 2023.

Morris H. Degroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. doi: https://doi.org/10.2307/2987588. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2987588`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

Timo Dimitriadis, Tilmann Gneiting, and Alexander I. Jordan. Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited. *Proceedings of the National Academy of Sciences*, 118 (8):e2016191118, February 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2016191118.

Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, December 2019. ISSN 0001-0782. doi: 10.1145/3359786.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16091–16101, June 2023.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

Ali Ghadiri, Maurice Pagnucco, and Yang Song. Xtranprune: explainability-aware transformer pruning for bias mitigation in dermatological disease classification. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part X*, volume 15010 of *Lecture Notes in Computer Science*, pp. 749–758. Springer, 2024. doi: 10.1007/978-3-031-72117-5_70. URL `https://doi.org/10.1007/978-3-031-72117-5_70`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proc. 2017 Int. Conf. Machine Learning*, pp. 1321–1330. PMLR, July 2017. URL `http://proceedings.mlr.press/v70/guo17a.html`. ISSN: 2640-3498.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations (ICLR)*, 2016a. arXiv preprint arXiv:1510.00149.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016b. URL `http://arxiv.org/abs/1510.00149`.

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 (CVPR)*, pp. 770–778, 2016.

Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. In Jérôme Lang (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 2234–2240. ijcai.org, 2018. doi: 10.24963/IJCAI.2018/309. URL https://doi.org/10.24963/ijcai.2018/309.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Zehao Huang and Naiyan Wang. Data-driven sparse structure selection for deep neural networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pp. 317–334. Springer, 2018. doi: 10.1007/978-3-030-01270-0_19. URL https://doi.org/10.1007/978-3-030-01270-0_19.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30*, pp. 5580–5590, Long Beach, CA, USA, 2017.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Learnable uncertainty under Laplace approximations. In *Proc. 37th Conf. Uncertainty in Artificial Intelligence*, pp. 344–353. PMLR, December 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report UTML TR 2009-003, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo de Menezes Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Snip: single-shot network pruning based on connection sensitivity. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=B1VZqjAcYX.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017a. URL https://openreview.net/forum?id=rJqFGTslg.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017b. URL https://openreview.net/forum?id=rJqFGTslg.

Zhu Liao, Victor Quétu, Van-Tam Nguyen, and Enzo Tartaglione. Can Unstructured Pruning Reduce the Depth in Deep Neural Networks? In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1394–1398, October 2023. doi: 10.1109/ICCVW60793.2023.00151.

Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Group fisher pruning for practical network compression. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7021–7032. PMLR, 2021. URL http://proceedings.mlr.press/v139/liu21ab.html.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2755–2763. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.298. URL https://doi.org/10.1109/ICCV.2017.298.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJlnB3C5Ym.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/44956951349095f74492a5471128a7e0-Abstract-Conference.html.

Huizi Mao, Song Han, Jeff Pool, Wenshuo Li, Xingyu Liu, Yu Wang, and William J. Dally. Exploring the regularity of sparse structure in convolutional neural networks. *arXiv preprint arXiv:1705.08922*, 2017. URL https://arxiv.org/abs/1705.08922.

Fanxu Meng, Hao Cheng, Ke Li, Huixiang Luo, Xiaowei Guo, Guangming Lu, and Xing Sun. Pruning filter in filter. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ccb1d45fb76f7c5a0bf619f979c6cf36-Abstract.html.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14014–14024, 2019.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

Pallavi Mitra, Gesina Schwalbe, and Nadja Klein. Investigating calibration and corruption robustness of post-hoc pruned perception cnns: An image classification benchmark study. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pp. 3542–3552. IEEE, 2024. doi: 10.1109/CVPRW63382.2024.00358. URL https://doi.org/10.1109/CVPRW63382.2024.00358.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 11264–11272. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01152.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-Wise Relevance Propagation: An Overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pp. 193–209. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_10.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), February 2015. ISSN 2374-3468. doi: 10.1609/aaai.v29i1.9602.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In Luc De Raedt and Stefan Wrobel (eds.), *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pp. 625–632. ACM, 2005. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE Conf. Comput. Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, volume 2, pp. 38–41. Computer Vision Foundation / IEEE, 2019.

Manuel Nonnenmacher, Thomas Pfeil, Ingo Steinwart, and David Reeb. SOSP: efficiently capturing global correlations by second-order structured pruning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=t5EmXZ3ZLR.

Jun-Hyung Park, Yeachan Kim, Junho Kim, Joon-Young Choi, and SangKeun Lee. Dynamic structure pruning for compressing cnns. *arXiv preprint arXiv:2303.09736*, 2023. URL https://arxiv.org/abs/2303.09736.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

Hassan Sajjad, Fadhl Eryani, Shengqi Sun, Rui Zhang, Yang Liu, Xipeng Qiu, and Holger Schwenk. A survey on pruning in transformer-based natural language processing models. *ACM Computing Surveys (CSUR)*, 55(12):1–37, 2023.

Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, January 2023. ISSN 1573-756X. doi: 10.1007/s10618-022-00867-8.

Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. HYDRA: Pruning Adversarially Robust Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19655–19666. Curran Associates, Inc., 2020.

Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3183–3193, Montréal, Canada, 2018. Curran Associates, Inc.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=PxoFut3dWW.

Yan Sun, Wenjun Xiong, and Faming Liang. Sparse deep learning: A new framework immune to local traps and miscalibration. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 22301–22312, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bb836c01cdc9120a9c984c525e4b1a4a-Abstract.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proc. 34th Int. Conf. Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, Sydney, NSW, Australia, 2017. PMLR.

Ji Tao, Xuxi Dai, Yuwei Wu, Xiaojie Jin, Jiashi Feng, and Ziwei Liu. Sparsevit: Accelerating vision transformers via holistic sparse attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16967–16977, 2023.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pp. 10347–10357. PMLR, 2021.

Juozas Vaicenavicius, David Widmann, Carl R. Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3459–3467. PMLR, 2019. URL `http://proceedings.mlr.press/v89/vaicenavicius19a.html`.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1580. URL `https://doi.org/10.18653/v1/p19-1580`.

Chaoqi Wang, Guodong Zhang, and Roger B. Grosse. Picking winning tickets before training by preserving gradient flow. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=SkgsACVKPH`.

Lihua Wang, Heinrich Jiang, Ofir Nachum, and Carlos Guestrin. Calibration for safe and reliable ai. *arXiv preprint arXiv:2301.12673*, 2023.

Ziheng Wang. Sparsert: Accelerating unstructured sparsity on gpus for deep learning inference. In *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2020. doi: 10.1145/3410463.3414654. URL `https://doi.org/10.1145/3410463.3414654`.

Oliver Willers, Sebastian Sudholt, Shervin Raafatnia, and Stephanie Abrecht. Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks. In António Casimiro, Frank Ortmeier, Erwin Schoitsch, Friedemann Bitsch, and Pedro Ferreira (eds.), *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, Lecture Notes in Computer Science, pp. 336–350, Cham, 2020. Springer International Publishing. ISBN 978-3-030-55583-2. doi: 10.1007/978-3-030-55583-2_25.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4697–4708. Curran Associates, Inc., 2020.

Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, and Jan Kautz. Global vision transformer pruning with hessian-aware saliency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18547–18557. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01779. URL `https://doi.org/10.1109/CVPR52729.2023.01779`.

Yazhou Yang and Marco Loog. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2646–2651, December 2016. doi: 10.1109/ICPR.2016.7900034.

Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition*, pp. 107899, February 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.107899.

Lu Yu and Wei Xiang. X-pruner: explainable pruning for vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24355–24363. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02333. URL `https://doi.org/10.1109/CVPR52729.2023.02333`.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Carla E. Brodley and Andrea Pohoreckyj Danyluk (eds.), *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pp. 609–616. Morgan Kaufmann, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pp. 694–699. ACM, 2002. doi: 10.1145/775047.775151. URL `https://doi.org/10.1145/775047.775151`.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. 2016 IEEE Conf. Comput. Vision and Pattern Recognition*, pp. 2921–2929, Las Vegas, NV, USA, 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.319.

A Appendix

You may include other additional sections here.