COMPUTER RESEARCH AND MODELING
2024 VOL. 16 NO. 4 P. 927–938
DOI: 10.20537/2076-7633-2024-16-4-927-938

UDC: 004.9

Image classification based on deep learning with automatic relevance determination and structured Bayesian pruning

Cong Thang Pham[1,a], Minh Nhat Phan[1], Thi Thu Thao Tran[2]

[1]The University of Danang — University of Science and Technology,
54 Nguyen Luong Bang st., Danang, 550000, Vietnam
[2]The University of Danang — University of Economics,
71 Ngu Hanh Son st., Danang, 550000, Vietnam

E-mail: [a] pcthang@dut.udn.vn (corresponding author)

Received 29.01.2024, after completion — 06.05.2024.
Accepted for publication 27.05.2024.

Deep learning's power stems from complex architectures; however, these can lead to overfitting, where models memorize training data and fail to generalize to unseen examples. This paper proposes a novel probabilistic approach to mitigate this issue. We introduce two key elements: Truncated Log-Uniform Prior and Truncated Log-Normal Variational Approximation, and Automatic Relevance Determination (ARD) with Bayesian Deep Neural Networks (BDNNs). Within the probabilistic framework, we employ a specially designed truncated log-uniform prior for noise. This prior acts as a regularizer, guiding the learning process towards simpler solutions and reducing overfitting. Additionally, a truncated log-normal variational approximation is used for efficient handling of the complex probability distributions inherent in deep learning models. ARD automatically identifies and removes irrelevant features or weights within a model. By integrating ARD with BDNNs, where weights have a probability distribution, we achieve a variational bound similar to the popular variational dropout technique. Dropout randomly drops neurons during training, encouraging the model not to rely heavily on any single feature. Our approach with ARD achieves similar benefits without the randomness of dropout, potentially leading to more stable training.

To evaluate our approach, we have tested the model on two datasets: the Canadian Institute For Advanced Research (CIFAR-10) for image classification and a dataset of Macroscopic Images of Wood, which is compiled from multiple macroscopic images of wood datasets. Our method is applied to established architectures like Visual Geometry Group (VGG) and Residual Network (ResNet). The results demonstrate significant improvements. The model reduced overfitting while maintaining, or even improving, the accuracy of the network's predictions on classification tasks. This validates the effectiveness of our approach in enhancing the performance and generalization capabilities of deep learning models.

Keywords: automatic relevance determination, Bayesian deep neural networks, truncated log-normal variational approximation, macroscopic image

Citation: *Computer Research and Modeling*, 2024, vol. 16, no. 4, pp. 927–938.

The authors express their gratitude to the anonymous reviewers for their insightful comments and suggestions. Pham Cong Thang, the corresponding author, would also like to thank his colleagues at the IT Faculty, DUT, Danang, Vietnam, for their helpful comments.

Introduction

In the field of image classification, Deep Neural Networks (DNNs) have made remarkable progress. However, when dealing with small datasets, neural networks tend to become overparameterized, leading to overfitting due to the abundance of weight parameters in the layers. To address this issue and mitigate overfitting during training, researchers have proposed various compression techniques to design lightweight models, including neural network pruning. Dropout regularization, initially suggested by Hinton et al. as a method to prevent overfitting [Hinton et al., 2012], involves randomly suppressing the activation levels of certain neurons during network training. This strategy enhances the model's feature extraction performance by retaining all neurons and recovering all training parameters. Zhou and Luo [Zhou, Luo, 2018] have proposed a strategy that considers the likelihood of node deletion based on the magnitude of the activation value. This approach enhances the model's feature extraction capabilities by selectively pruning nodes, thereby improving overall efficiency.

In addition to overfitting, compression and acceleration present significant challenges for neural networks, especially when their performance heavily depends on model parameters and computational cost. Further investigation into variational dropout reveals the possibility of reducing the size of the original network design and creating an extremely sparse model by examining individual dropout rates for each weight [Molchanov, Ashukha, Vetrov, 2017].

While general sparsity offers a means of neural network compression, most present-day deep neural network (DNN)-oriented software struggles to effectively handle sparse matrices, potentially maintaining the same assessment time for network evaluation. Simultaneously, enforcing organized sparsity in data tensors or convolutional filters can lead to acceleration. Structured sparsity techniques vary from the straightforward removal of redundant neurons or adoption of convolutional filters to more intricate patterns, such as those implemented in Perforated CNNs for eliminating redundant rows in intermediate dataframe matrices during convolutions [Figurnov et al., 2016]. Group-wise Brain Damage utilizes group-wise sparsity in convolutional filters [Lebedev, Lempitsky, 2016], and Structured Sparsity Learning offers a mechanism for selectively pruning entire convolutional filters or layers within residual networks [Wen et al., 2016]. These approaches facilitate the creation of parsimonious models while preserving accuracy, providing realistic acceleration with minimal adjustments to the existing model.

A meticulously adapted log-uniform prior, as proposed by Molchanov et al. [Molchanov, Ashukha, Vetrov, 2017; Kingma, Salimans, Welling, 2015], promotes sparsity while preserving model correctness and overcoming the drawbacks associated with improper priors [Hron, Matthews, Ghahramani, 2018]. Going beyond its conventional application, dropout can be viewed through the lens of variational inference and automatic relevance determination [Mackay, 1995; Neal, 2012], offering an alternative interpretation that mitigates its known deficiencies.

In this paper, these two methods are combined to create a model with substantial acceleration in convolutional neural networks, demonstrating an acceptable level of accuracy loss and a high level of group sparsity. The proposed model is designed based on VGG and ResNet architectures using CIFAR-10 and a Macroscopic Image of Wood dataset.

Related work

Pruning techniques in Bayesian neural network

In [Louizos, Ullrich, Welling, 2017], the pruning of neurons and all their incoming and outgoing weights is achieved through Bayesian modeling using sparsity-inducing priors on hidden units, instead of individual weights. In contrast, the author in [Neklyudov et al., 2017] primarily applies the Bayesian model to eliminate certain channels in each layer of the network for pruning purposes. The authors in [Zhou et al., 2019] achieved more optimal pruning by accounting for interlayer dependencies during the pruning process and introduced a novel dropout-based redundancy measure to compute posterior hypothesis inter-layer relationships. Despite their differences, applying these techniques to the Bayesian model ultimately optimizes a weight importance metric that directly influences the pruning process.

Variational Dropout, initially proposed by Kingma et al. in 2015 [Kingma, Salimans, Welling, 2015], investigates the relationship between dropout and stochastic gradient variational Bayesian inference. Subsequent work by Molchanov et al. [Molchanov, Ashukha, Vetrov, 2017] further explores Variational Dropout, demonstrating its utility as a technique for model compression. This approach enables neural networks to utilize significantly fewer parameters while maintaining a surprisingly high level of accuracy.

Gaussian Dropout, as suggested in [Srivastava et al., 2014], represents a local reparameterization of stochastic gradient variational Bayesian (SGVB). The goal is to treat the dropout ratio as a parameter that can be learned during training, rather than as a fixed hyperparameter adjusted by the user.

In 2016, Gal and Ghahramani introduced the concept of Monte Carlo Dropout [Gal, Ghahramani, 2016], representing a deep Gaussian process in Bayesian approximation. A deep Gaussian process generates a probability distribution as its output, and the parameters of this distribution are estimated during testing using conventional dropout. This approach forms the basis for subsequent research [Gal, 2016; Lakshminarayanan, Pritzel, Blundell, 2017; Zhu, Laptev, 2017; Jungo et al., 2018], which focuses on characterizing the uncertainty of model outputs to address expressing uncertainty in deep learning without compromising test accuracy.

The idea that various neurons in a neural network should have varied dropout probabilities based on their perceived significance for specific characteristics was proposed by Li et al. through Evolutionary Dropout [Li, Gong, Yang, 2016]. Applied to both shallow and deep neural networks, Evolutionary Dropout determines the dropout ratio in shallow networks using the second-order statistics of input data features. In deep networks, the dropout ratio for each layer is determined in real-time using the output of that layer for each batch. Evolutionary Dropout significantly accelerates convergence while maintaining better accuracy results compared to ordinary dropout.

Wood identification

There are currently two primary techniques used for wood species identification: automatic and manual. Identifying the species of wood using the manual method is the conventional and most popular approach. This technique is based on the morphological (size, shape, color, and wood grain), mechanical (hardness, strength, and ductility), and chemical (chemical composition) characteristics of wood. Nevertheless, this method requires a high level of expertise and specialized wood knowledge from the practitioner [Silva, Bordalo, Pissarra, 2020]. In contrast, the automatic deep learning model for wood classification offers benefits such as reliability, accuracy, and lower requirements for practitioner experience [Silva, Bordalo, Pissarra, 2020].

Utilizing a diverse dataset containing 77 Congolese wood species, the authors in [Rosa et al., 2022] employed the multi-view random forest (MVRF) model to achieve accurate species identification. This innovative approach incorporated information from three distinct anatomical planes (cross-sectional, tangential, and radial), ensuring a comprehensive understanding of each wood sample [Rosa et al., 2022]. For wood species identification, integrating both cross-sectional and tangential features in the multi-view random forest model resulted in a substantial improvement in accuracy compared to using only cross-sectional features. While the inclusion of the radial plane also provided a slight performance gain, it was not as significant as the combined effect of cross-sectional and tangential features. Unlike traditional methods that rely solely on local phase quantization (LPQ) features, the MVRF model emerged as a clear winner, achieving significantly superior identification accuracy. This success can be attributed to the additional information harnessed by the three-plane analysis, enriching the model's understanding of wood anatomy. Acknowledging the complexity of the wood species identification task and potential biases of k-folds, the authors opted for the leave-k-tree-out method for a more rigorous and reliable evaluation of their models, ensuring generalizability beyond the training data.

In differentiating between Juniperus cedrus and J. phoenicea var. canariensis, the authors of [Esteban et al., 2009] employed a feedforward multilayer perceptron (MLP) network, a type of artificial neural network, achieving a remarkable 92% accuracy. However, a limitation of this study is the lack of data diversity, as the performance of this model significantly decreases when the dataset is expanded [Silva et al., 2022]. The authors of [Lens et al., 2020] applied the ResNet101 model combined with an SVM classifier to identify 112 tree species, mainly in the tropics. This model focuses on microscopic analysis rather than macroscopic analysis and achieved an identification accuracy of 95.6 % using cross-sections. The authors of [Filho et al., 2014] explored machine learning methods for wood species identification, proposing a two-level divide-and-conquer strategy for classification. Experiments on the Forest Species Database — Macroscopic dataset of 41 species yielded a result of 97.07 %. The authors of [He et al., 2021] proposed a composite model from three convolutional neural network (CNN) models, using transfer learning to increase training speed. This method achieved an accuracy of 98.81 % after training on a dataset of 41 wood species and 11,984 images. The results also highlighted the efficacy of using macroscopic images for high-performance identification, as opposed to microscopic images. However, both of these studies observed a decrease in accuracy when the dataset was expanded [Silva et al., 2022].

Methodology

Variational automatic relevance determination for neural network

At the core of our analysis are the weight matrices $W = \left(W^{(l)}\right)_{l=1}^{L}$ present in L layers, examined within a data collection $\mathcal{D}$ consisting of paired input and output elements $(X, Y) = (x_n, y_n)_{n=1}^{N}$. Assuming the existence of a parametric model $p(\mathcal{D} \mid W) = \prod_{i=1}^{N} p(y_i \mid x_i, W)$, where $W \in \mathbb{R}^D$ establishes a relation between x and y governed by parameters W. Hyperparameters $\tau \in \mathbb{R}^H$ act as levers, adjusting the prior distribution $p(W \mid \tau)$ to influence the possible values of parameters W. In Bayesian neural networks (BNNs), addressing the intractability of the posterior $p(W \mid \mathcal{D}, \tau) =$ $= p(\mathcal{D} \mid W)p(W \mid \tau)/p(\mathcal{D} \mid \tau)$ involves employing a tractable parametric prior $p(W \mid \tau)$. We strategically optimize its hyperparameters τ to maximize the dataset's marginal likelihood (evidence), effectively approximating the posterior [Kharitonov, Molchanov, Vetrov, 2018]:

$$\tau^* = \arg\max_{\tau} p(\mathcal{D} \mid \tau) = \arg\max_{\tau} \int p(\mathcal{D} \mid W)p(W \mid \tau)dW. \tag{1}$$

Doubly Stochastic Variational Inference (DSVI) [Titsias, Lazaro-Gredilla, 2014] provides a method for approximating intractable posterior distributions in (1) by strategically designing an approximate posterior $q(W \mid \phi)$ and minimizing the KL-divergence $D_{KL}(q(W \mid \phi), \|, p(W \mid \tau))$. This process effectively translates to optimizing the Evidence Lower Bound (ELBO) by considering both parameters ϕ and hyperparameters τ, as follows:

$$\log p(\mathcal{D} \mid \tau) \geqslant \mathcal{L}(\phi, \tau) = \mathbb{E}_{q(W|\phi)}[\log p(\mathcal{D} \mid W)] - D_{KL}(q(W \mid \phi) \parallel p(W \mid \tau)) \to \max_{\phi,\tau}. \tag{2}$$

Incorporating an Automatic Relevance Determination (ARD) prior [Mackay, 1995; Neal, 2012] given by $p(W \mid \tau) = \prod_{i=1}^{D} \mathcal{N}\left(W_i \mid 0, \tau_i^{-1}\right)$ and leveraging $q(W \mid \mu, \sigma) = \prod_{i=1}^{D} \mathcal{N}\left(W_i \mid \mu_i, \sigma_i^2\right)$, we can analytically derive the optimal hyperparameter values as $\tau_i^* = \left(\mu_i^2 + \sigma_i^2\right)^{-1}$ [Titsias, Lazaro-Gredilla, 2014]. This leads to the simplified Evidence Lower Bound (ELBO) expression of (2) as follows [Titsias, Lazaro-Gredilla, 2014]:

$$\mathcal{L}_{ARD}(\mu, \sigma) = \sum_{i=1}^{N} \mathbb{E}_{q(W|\mu,\sigma)}[\log p(y_i \mid x_i, W)] - \frac{1}{2}\sum_{j=1}^{D} \log\left(1 + \frac{\mu_j^2}{\sigma_j^2}\right) = \mathcal{L}_D(\mu, \sigma) + \mathcal{R}_{ARD}(\mu, \sigma) \to \max_{\mu,\sigma}. \tag{3}$$

The local reparameterization trick [Kingma, Salimans, Welling, 2015] serves as a crucial variance reduction technique, facilitating efficient optimization of this estimation through stochastic gradient methods. Under the restricted variational approximation $q(W \mid \mu) = \prod_{i=1}^{\mathcal{D}} \mathcal{N}\left(W_i \mid \mu_i, \alpha_i \mu_i^2\right)$, where $\alpha_i > 0$, the Automatic Relevance Determination (ARD) objective (3) simplifies significantly. The regularizer term, being constant, no longer impacts optimization. This results in a streamlined objective function that depends solely on the parameters μ. Mathematically, it can be expressed as

$$\mathcal{L}_{ARD}(\mu) = \mathcal{L}_D(\mu) = \sum_{i=1}^{N} \mathbb{E}_{q(W|\mu)}[\log p(y_i \mid x_i, W)] \to \max_{\mu}. \tag{4}$$

Connection with Log-Normal Multiplicative Noise Pruning

Log-Normal Multiplicative Noise Pruning is a technique that achieves organized sparsity within neural networks through a unique approach. This innovative mechanism fosters a structured pruning process, enabling the removal of entire neurons or convolutional channels for a leaner, more efficient network architecture. It introduces a distinct dropout-like layer, accompanied by a targeted regularization term, to strategically inject multiplicative noise θ into the outputs x with I features from preceding layers:

$$y_i = x_i \cdot \theta_i, \tag{5}$$

$$\theta_i \sim p_{noise}(\theta_i). \tag{6}$$

A fully-factorized improper log-uniform prior, denoted as $\mathrm{LogU}_\infty(\cdot)$, is selected for the noise θ with the prior distribution $p(\theta)$, ensuring model sparsity and possessing an infinite domain. This distribution operates effectively for deep neural networks because of its sparsification properties [Molchanov, Ashukha, Vetrov, 2017]. With $\theta_i > 0$, $p(\theta) = \prod_{i=1}^{I} p(\theta_i)$ in (5) and (6), the log-uniform prior for sparsity is expressed as follows:

$$p(\theta_i) = \mathrm{LogU}_\infty(\theta_i) \propto \frac{1}{\theta_i}.$$

In order to perform variational inference, an approximation family $q(\theta \mid \phi)$ needs to be determined for the posterior distribution $p(\theta \mid \mathcal{D}) \approx q(\theta \mid \phi)$. This variational distribution is chosen to be a fully-factorized log-normal distribution, with $\varphi_i = \left(\mu_i, \sigma_i^2\right)$:

$$q(\theta \mid \phi) = \prod_{i=1}^{I} q(\theta_i \mid \mu_i, \sigma_i) = \prod_{i=1}^{I} \mathrm{LogN}(\theta_i \mid \varphi_i),$$

$$\theta_i \sim \mathrm{LogN}(\theta_i \mid \varphi_i) \Leftrightarrow \log\theta_i \sim \mathcal{N}(\log\theta_i \mid \varphi_i).$$

While the log-normal posterior and log-uniform prior offer enticing advantages, their combination for variational lower bound maximization presents optimization challenges. The log-uniform prior's improper nature leads to infinite KL-divergence between the log-normal distribution $\mathrm{LogN}(\varphi)$ and a log-uniform distribution LogU_∞ for finite parameter φ values, rendering the optimization ill-posed. To resolve this problem, a proper probabilistic model is obtained with the prior and the posterior:

$$\begin{gathered}
p(\theta_i) = \mathrm{LogU}_\infty(\theta_i) \Rightarrow \mathrm{LogU}_{[a,b]}(\theta_i),\\
q(\theta_i) = \mathrm{LogN}(\theta_i \mid \varphi_i) \Rightarrow \mathrm{LogN}_{[a,b]}(\theta_i \mid \varphi_i),\\
\mathrm{LogP}_{[a,b]}(\theta_i) \propto \mathrm{LogP}_\infty(\theta_i) \cdot \mathrm{I}_{[a,b]}(\log \theta_i).
\end{gathered}$$

Leveraging the defined notation, where $\mathcal{L}_{ARD}(\mu)$ (4) signifies the ARD model loss, q and p represent truncated distributions, and $\phi(\cdot)$ and $\Phi(\cdot)$ symbolize the standard normal's PDF and CDF, the final loss is computed as follows:

$$\mathcal{L} = \mathcal{L}_{ARD}(\mu) - \sum_{l=1}^{L} KL(q(\theta_l \mid \mu_l, \sigma_l) \parallel p(\theta_l)),$$

where

$$KL(q(\theta_l \mid \mu_l, \sigma_l) \parallel p(\theta_l)) = \sum_{i=1}^{I} KL(q(\theta_{l,i} \mid \mu_{l,i}, \sigma_{l,i}) \parallel p(\theta_{l,i})),$$

$$KL(q(\theta_{l,i} \mid \mu_{l,i}, \sigma_{l,i}) \parallel p(\theta_{l,i})) = \log \frac{b-a}{\sqrt{2\pi e \sigma_{l,i}^2}} - \log(\Phi(\beta_{l,i}) - \Phi(\alpha_{l,i})) - \frac{\alpha_{l,i}\phi(\alpha_{l,i}) - \beta_{l,i}\phi(\beta_{l,i})}{2(\Phi(\beta_{l,i}) - \Phi(\alpha_{l,i}))},$$

$$\alpha_{l,i} = \frac{a-\mu_{l,i}}{\sigma_{l,i}}, \quad \beta_{l,i} = \frac{b-\mu_{l,i}}{\sigma_{l,i}}.$$

Experimental results

The performance of the proposed method is assessed on two datasets: the standard CIFAR-10 benchmark and a real-world dataset of macroscopic wood images.

Dataset

Four individual datasets were integrated to create the comprehensive dataset used in this study: the Tropical Forest Species[1], the Brazilian Flora Species[2], the Wood Species Dataset[3], and the Forest Species Database[4]. This integration was aimed at broadening the diversity for classification. During data preparation, images from classes with identical names were merged. Additionally, naming errors introduced by collectors were addressed by verifying and correcting the accurate species names, leading to the consolidation of corresponding image collections. The final dataset comprises 16,346 images from 75 species.

The combined dataset and detailed statistical information about the data sources are presented in Table (1) and Fig. (1). The distribution of image counts for each wood type in the combined dataset is depicted in Fig. (2). As indicated in Table (1) and Fig. (2), each species is represented by an average of 218 images, with a range from 24 to 1332 images per species. The Tropical Forest Species source contributes over 65 % of the dataset's images, whereas the Wood Species Dataset accounts for only about 4 % of the total.

[1] https://web.inf.ufpr.br/vri/databases/forest-species-database-macroscopic (accessed 18-03-2023).

[2] https://zenodo.org/record/2545611 (accessed 18-03-2023).

[3] https://data.mendeley.com/datasets/yzzcbyvgmh/3 (accessed 18-03-2023).

[4] https://data.mendeley.com/datasets/cc78ftcdf9/1 (accessed 18-03-2023).

Table 1. Statistical analysis of dataset

Dataset	Number of images	Percentage	Number of species	Average
Forest Species Database	2942	17.99 %	41	71.76
Wood Species Dataset	709	4.34 %	12	59.08
Tropical Forest Species	10,795	66.04 %	11	981.36
Brazillian	1901	11.63 %	46	41.32
Total	16,346	100 %	75	217.95

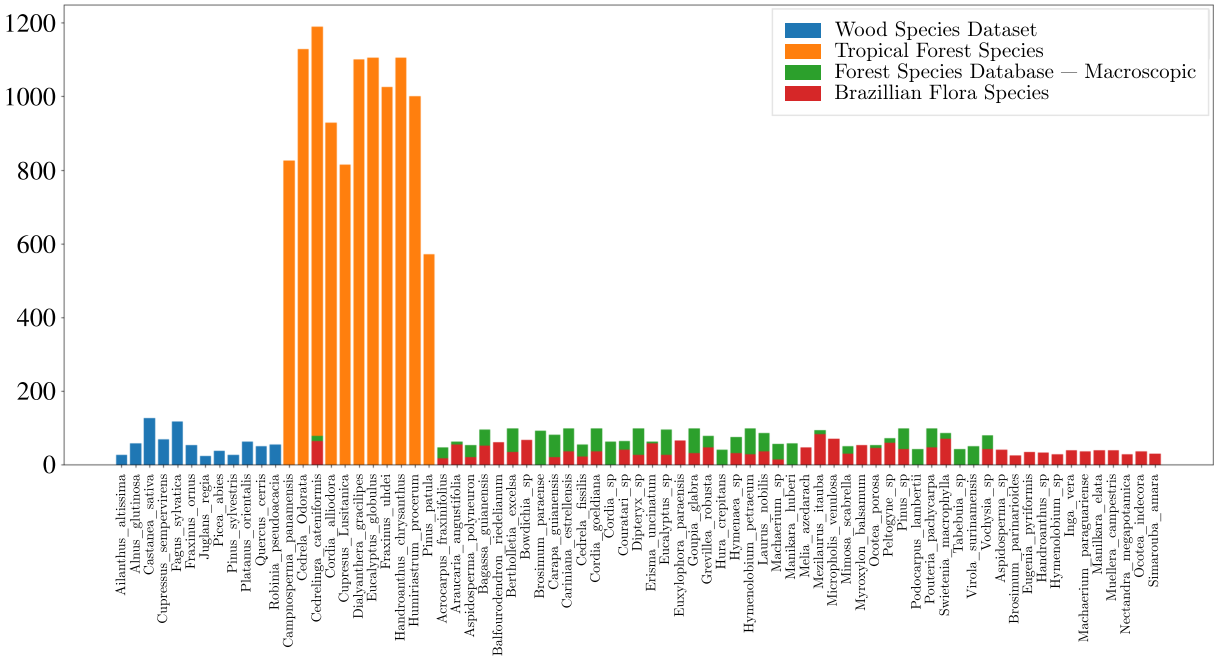

Figure 1. Number of images per species in each datasets

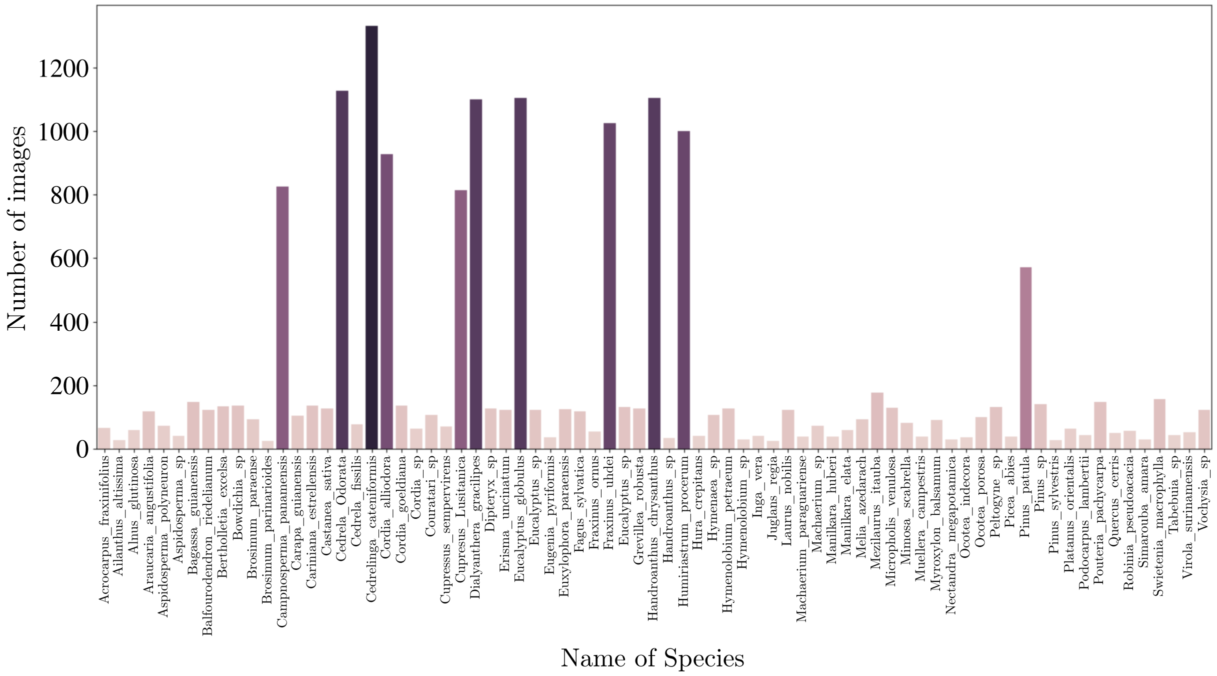

Figure 2. Number of images per species in the total dataset

Experiment setup

We built models based on the VGG-19 [Simonyan, Zisserman, 2014] and ResNet-18 [He et al., 2016] architectures and trained them on the CIFAR-10 and Macroscopic Image of Wood datasets for 300 epochs, with a batch size of 2048. We selected Adam as our optimizer, starting with a learning rate of 10^{-3} and employing Cosine Annealing for a smooth decay to 10^{-5} by the end of training. To prevent underfitting, we carefully scaled the regularizer terms by 0.05 and fine-tuned the truncation parameters a and b to ensure model regularization without amplifying input signals. We constrained θ to the range [0, 1], setting the right truncation threshold b to 0, as empirical evidence shows minimal impact from the left truncation parameter a, which was set at $a = -20$ consistently throughout the experiments. The proposed method was implemented with the optimal parameter configuration described in the original study [Neklyudov et al., 2017] for comparative analysis. We divided the Macroscopic Image of Wood dataset into training, validation, and testing sets, allocating 60 %, 20 %, and 20 % of the data, respectively. The experiments were conducted using Python (v.3.9) and PyTorch (v.2.1) as the computational backbone, with the NVIDIA GeForce RTX 3090 graphics card, which features a 24 Gb memory capacity.

Evaluation

On CIFAR-10

Table 2. Evaluation on CIFAR-10 testing set

Model	Accuracy	Sparsity
Pruning ARD ResNet-18	96.352	0.931
Pruning ARD VGG-19	95.048	0.932

In Table 2, we report accuracy and sparsity for ResNet-18 and VGG-19 trained with Automatic Relevance Determination (ARD) combined with Structured Bayesian Pruning. Figure 3 depicts a comparison between the accuracy of the original model and the model trained with ARD combined with Structured Bayesian Pruning. It is evident that the proposed network model exhibits high sparsity with negligible degradation in classification accuracy. Specifically, ResNet-18 and VGG-19, trained with ARD combined with Structured Bayesian Pruning, achieve accuracies of 96.352 % and 95.048 %, respectively. Furthermore, the accuracy of the suggested models decreases by only 2–3 % compared to the original models. Both models, built with the new technique, exhibit more than 93 % sparsity. Higher sparsity helps networks reduce numbers of active connections by eliminating connections entirely, the model bypassed the need to explicitly multiply each input. This eliminates unnecessary computations, leading to improved efficiency. In Bayesian statistics, hierarchical priors are a way to introduce prior information not just about the model parameters, but also about the hyperparameters themselves. The approach also accommodates the application of hierarchical priors to hyperparameters, providing a mechanism for optimizing the sparsity-accuracy trade-off. Notably, the original model exhibits an inadequate level of generalization on the test set despite satisfactory performance on the training set. In contrast, the suggested model mitigates overfitting issues. Within the first 50 epochs, the proposed model demonstrates notable acceleration in achieving accurate predictions, transitioning to a slower approach and ultimately reaching optimal performance. The ARD model attains a training accuracy of over 80 % after just 30 epochs, outperforming other models that require a more extended training period to reach similar accuracy levels. Overall, these findings suggest that the ARD technique effectively improves the accuracy of deep learning models. ARD models appear less prone to overfitting than other models, achieving high accuracy on both training and test datasets.

The impressive complexity-learning capabilities of large networks are undermined by their vulnerability to overfitting and their unsuitability for devices lacking computational muscle. Pruning,

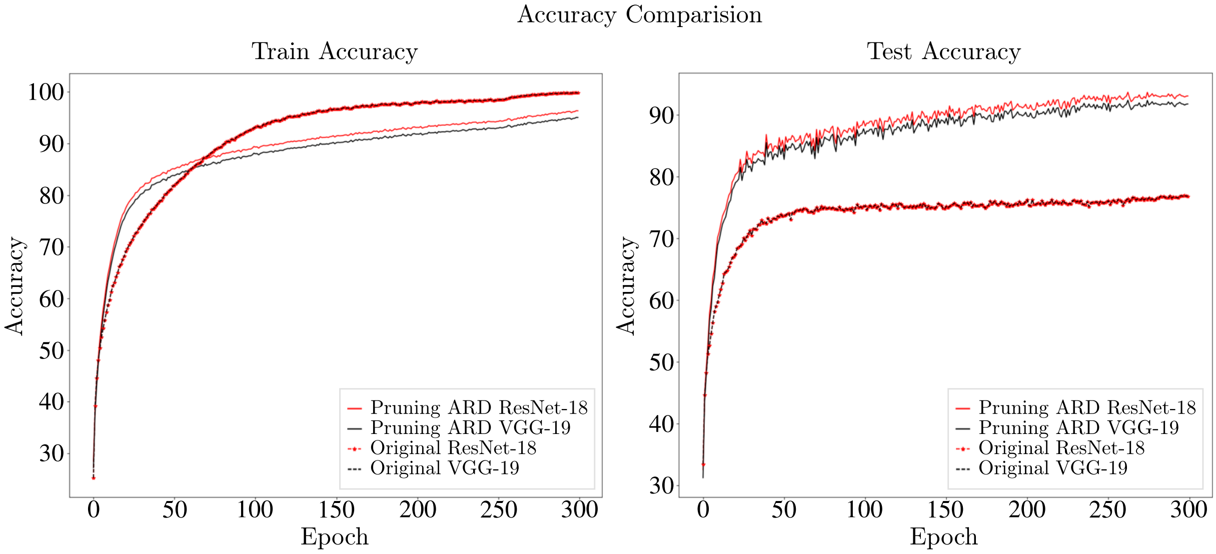

Figure 3. Training and Testing Accuracy after each epoch on CIFAR-10 dataset

a technique that optimizes network size with minimal impact on learning complex functions, serves as a bridge between powerful yet resource-intensive networks and efficient but limited models. Specifically addressing the issue of Binary Neural Network (BNN) complexity, our proposed pruning method takes a significant step forward by successfully eliminating a substantial portion of the network parameters. However, the current approach to structured pruning, though successful in shrinking the model, carries the risk of generating sparse structures that may negate potential computational benefits. To address this concern, our suggested method outperforms other studies [Mathew, Rowe, 2023; Beckers et al., 2023] in terms of accuracy and sparsity, offering better results. This method represents a promising advancement in overcoming the challenges associated with pruning BNNs and enhancing their efficiency.

On Macroscopic Image of Wood

Table 3. Evaluation on Macroscopic Image of Wood testset

Model	Accuracy	F1 Score	Precision	Recall	Sparsity
Pruning ARD ResNet-18	91.876	0.901	0.903	0.923	0.933
Original ResNet-18	91.572	0.913	0.907	0.919	0.226
Pruning ARD VGG-19	90.538	0.889	0.892	0.887	0.932
Original VGG-19	90.782	0.899	0.897	0.901	0.235

Table 3 and Fig. 4 present information on four metrics related to the training process of four models. These metrics encompass accuracy, F1 Score, Precision, and Recall. In order to assess model performance on imbalanced datasets, macro-averaging is relied upon for precision, recall, and F1-score. This technique is particularly effective because it considers the size of each class, ensuring a fair evaluation even when some classes have far fewer samples [Mortaz, 2020]. While accuracy might seem like a straightforward metric, it can be misleading for imbalanced data. A model could simply predict the majority class all the time and achieve high accuracy, yet completely miss the mark on the minority classes that are often more important. Accuracy is therefore employed for visualization purposes in Fig. 4. It serves to illustrate general trends, such as how the model's overall performance changes with increasing class imbalance in the training data. Additionally, accuracy's simplicity makes it an accessible way to convey the results. As observed, the suggested models achieve high accuracy

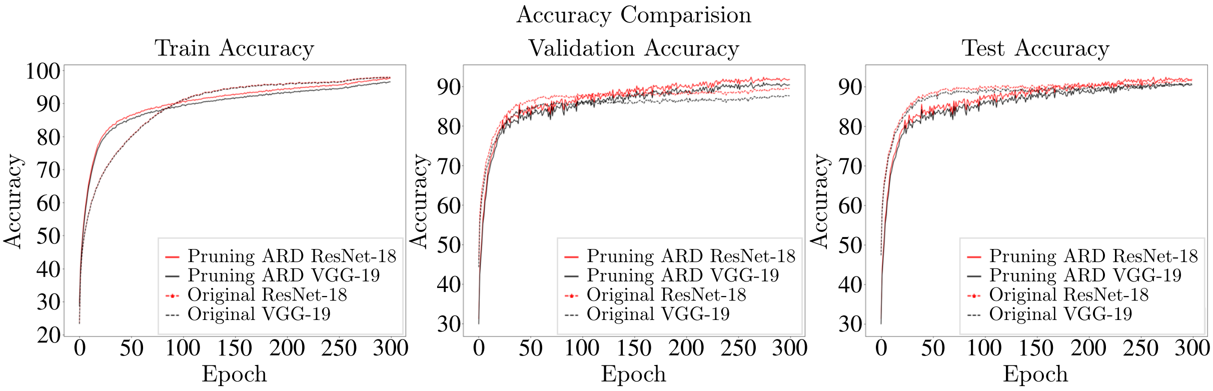

Figure 4. Training, Validation and Testing Accuracy after each epoch on Macroscopic Image of Wood dataset

without any reduction in these metrics when compared to the original model. The Automatic Relevance Determination (ARD) model demonstrates rapid convergence, surpassing a training accuracy of 80 % within 50 epochs. This swift learning pace highlights its efficient learning capabilities, outperforming other models. While the test accuracy gap narrows slightly compared to training, ARD maintains its advantage, showcasing superior generalizability. In comparison to the original model, the proposed method achieves a substantial 4 times compression in sparsity. This implies that the ARD model has four times as many zeros for the same number of parameters, leading to a fourfold boost in calculation speed due to efficient zero-computation. This significant improvement in efficiency surpasses the original model's performance.

Despite achieving high accuracy in wood classification based on macroscopic images, the proposed model falls short of previous studies [Esteban et al., 2009; Lens et al., 2020], due to data imbalance and data expansion. To enhance its performance and potentially surpass previous studies, future work will address data imbalance using techniques such as Focal Loss or Upsample. This approach holds promise for further improvements in accuracy and metrics, including F1 score, precision, and recall.

Conclusion

This work introduces a groundbreaking approach that combines variational Automatic Relevance Determination (ARD) approximation with fully factorized Gaussian variational posterior distributions, along with a dropout-like layer for noise introduction. Its iterative pruning mechanism stands out, and enables the model to achieve sparsity of over 90 % while maintaining as accuracy loss of less than 1–2 %. Demonstrably effective on VGG and ResNet, the flexibility of this technique extends to a wider realm of Convolutional Neural Network (CNN) architectures. Its impact is poised to reshape the future of CNN design, paving the way for architectures like MobileNet, ConvNet, and ConvNeXt to leverage its power. In this study, we construct the architecture of deep learning models and compare the effectiveness of this model to the original models that achieved high accuracy in wood classification based on macroscopic images. The ResNet-ARD model, designed to identify wood species by analyzing visual images, achieves an accuracy of 91.876 % on the test set, demonstrating the superiority of the model over its predecessors. The ARD technique enables the creation of models with fewer nonzero elements, reducing storage requirements, and potentially improving computational efficiency, all while maintaining the same level of accuracy.

References

Beckers J., Erp B. V., Zhao Z., Kondrashov K., Vries B. D. Principled pruning of Bayesian neural networks through variational free energy minimization. // IEEE Open Journal of Signal Processing. — 2023. — Vol. 5. — P. 195–203.

Esteban L. G., Fernandez F. G., et al. Artificial neural networks in wood identification: the case of two Juniperus species from the Canary Islands // IAWA Journal. — 2009. — Vol. 30. — P. 87–94.

Figurnov M., Ibraimova A., Vetrov D., Kohli P. Perforated CNNs: acceleration through elimination of redundant convolutions // International Conference on Neural Information Processing Systems. — 2016. — P. 955–963.

Filho P. L. P., Oliveira L. S., Nisgoski S., Britto A. S. Forest species recognition using macroscopic images // Machine Vision and Applications. — 2014. — Vol. 25. — P. 1019–1031.

Gal Y. Uncertainty in deep learning. — University of Cambridge, 2016. — 174 p.

Gal Y., Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning // International Conference on Machine Learning, ICML. — 2016. — Vol. 48. — P. 1050–1059.

He T., Mu S., Zhou H., Hu J. Wood species identification based on an ensemble of deep convolution neural networks // Wood Research. — 2021. — Vol. 66. — P. 1–14.

He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // IEEE conference on computer vision and pattern Recognition (CVPR). — 2016. — P. 770–778.

Hinton G., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. Improving neural networks by preventing co-adaptation of feature detectors // CoRR. — 2012. — P. 1–18.

Hron J., Matthews A., Ghahramani Z. Variational Bayesian dropout: pitfalls and fixes // International Conference on Machine Learning. — 2018. — P. 2019–2028.

Jungo A. et al. Towards uncertainty-assisted brain tumor segmentation and survival prediction // International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI. BrainLes 2017. — 2018. — P. 474–485.

Kharitonov V., Molchanov D., Vetrov D. Variational dropout via empirical Bayes // eprint arXiv:1811.00596. — 2018. — P. 1–5.

Kingma D. P., Salimans T., Welling M. Variational dropout and the local reparameterization trick // International Conference on Neural Information Processing Systems. — 2015. — P. 2575–2583.

Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles // International Conference on Neural Information Processing Systems. — 2017. — P. 6405–6416.

Lebedev V., Lempitsky V. Fast convnets using group-wise brain damage // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2016. — P. 2554–2564.

Lens F., Liang C., Guo Y., et al. Computer-assisted timber identification based on features extracted from microscopic wood sections // IAWA Journal. — 2020. — Vol. 41. — P. 660–680.

Li Z., Gong B., Yang T. Improved dropout for shallow and deep learning // International Conference on Neural Information Processing Systems. — 2016. — P. 2531–2539.

Louizos C., Ullrich K., Welling M. Bayesian compression for deep learning // International Conference on Neural Information Processing Systems. — 2017. — P. 3290–3300.

MacKay D. J. C. Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks // Network: Computation in Neural Systems. — 1995. — Vol. 6. — P. 469–505.

Mathew S., Rowe D. B. Pruning a neural network using Bayesian inference // eprint arXiv:2308.02451. — 2023. — P. 1–26.

Molchanov D., Ashukha A., Vetrov D. Variational dropout sparsifies deep neural networks // International Conference on Machine Learning. — 2017. — Vol. 70. — P. 2498–2507.

Mortaz E. Imbalance accuracy metric for model selection in multi-class imbalance classification problems // Knowledge-Based Systems. — 2020. — Vol. 210. — P. 1–8.

Neal R. M. Bayesian learning for neural networks. — Lecture Notes in Computer Science. — 2012. — Vol. 118. — 204 p.

Neklyudov K., Molchanov D., Ashukha A., Vetrov P. Structured Bayesian pruning via log-normal multiplicative noise // International Conference on Neural Information Processing Systems. — 2017. — P. 6778–6787.

Rosa da Silva N., Deklerck V., Baetens J. M., et al. Improved wood species identification based on multi-view imagery of the three anatomical planes // Plant Methods. — 2022. — Vol. 18. — P. 1–17.

Silva J. L., Bordalo R., Pissarra J. Wood identification: an overview of current and past methods // ECR — Studies in Conservation & Restoration. — 2020. — Vol. 12. — P. 45–68.

Silva J. L., Bordalo R., Pissarra J., Palacios P. Computer vision-based wood identification: A review // Forests. — 2022. — Vol. 13. — P. 1–26.

Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // International Conference on Learning Representations, ICLR. — 2015. — P. 1–14.

Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting // The Journal of Machine Learning Research. — 2014. — Vol. 15. — P. 1929–1958.

Titsias M. K., Lazaro-Gredilla M. Doubly stochastic variational Bayes for non-conjugate inference // International Conference on Machine Learning, ICML. — 2014. — Vol. 32. — P. 1971–1979.

Wen W., Wu C., Wang Y., Chen Y., Li H. Learning structured sparsity in deep neural networks // International Conference on Neural Information Processing Systems. — 2016. — P. 2082–2090.

Zhou A., Luo K. Sparse dropout regularization method for convolutional neural networks // Journal of Chinese Computer Systems. — 2018. — Vol. 39. — P. 1674–1679.

Zhou Y., Zhang Y., Wang Y., Tian Q. Accelerate CNN via Recursive Bayesian Pruning // IEEE International Conference on Computer Vision, ICCV. — 2019. — P. 3305–3314.

Zhu L., Laptev N. Deep and confident prediction for time series at uber // IEEE International Conference on Data Mining, ICDM. — 2017. — P. 103–110.