SELECT HIGH-LEVEL FEATURES: EFFICIENT EXPERTS FROM A HIERARCHICAL CLASSIFICATION NETWORK

André Kelm, Niels Hannemann*, Bruno Heberle*, Lucas Schmidt*, Tim Rolff, Christian Wilms, Ehsan Yaghoubi, Simone Frintrop
Department of Computer Vision
University of Hamburg
Hamburg, Germany
andre.kelm@, 0hannema@informatik., bruno.heberle@, lucas.schmidt-1@studium., tim.rolff@, christian.wilms@, ehsan.yaghoubi@, simone.frintrop@uni-hamburg.de

ABSTRACT

This study introduces a novel expert generation method that dynamically reduces task and computational complexity without compromising predictive performance. It is based on a new hierarchical classification network topology that combines sequential processing of generic low-level features with parallelism and nesting of high-level features. This structure allows for the innovative extraction technique: the ability to select only high-level features of task-relevant categories. In certain cases, it is possible to skip almost all unneeded high-level features, which can significantly reduce the inference cost and is highly beneficial in resource-constrained conditions. We believe this method paves the way for future network designs that are lightweight and adaptable, making them suitable for a wide range of applications, from compact edge devices to large-scale clouds. In terms of dynamic inference our methodology can achieve an exclusion of up to 88.7 % of parameters and 73.4 % fewer giga-multiply accumulate (GMAC) operations, analysis against comparative baselines showing an average reduction of 47.6 % in parameters and 5.8 % in GMACs across the cases we evaluated.

1 INTRODUCTION

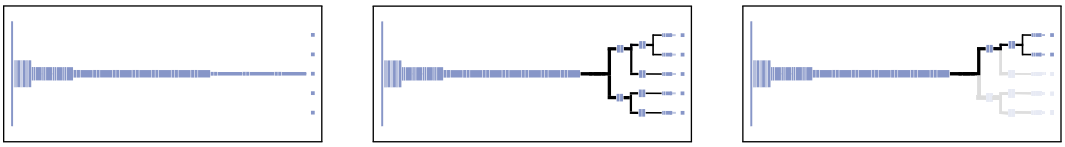
Wortsman et al. (2022) have shown that one way to achieve performance advantages in various deep learning (DL) tasks is to use larger models and more training material. High computational power and multi-stage training processes (Woo et al., 2023) create top performance, but typically do not fit easily on edge devices with limited resources. Cheng et al. (2023) summarize strategies such as model pruning, quantization, dynamic inference costs, and efficient architectures, highlighting their efficiency, but also noting the challenge of maintaining full performance.

Our method is intended to be orthogonal to these methods and offers the possibility of reducing the computational complexity without compromising predictive performance. Unlike other efficient methods, where efficiency is often learned within the optimization process, our approach can be controlled extrinsically to skip task-irrelevant high-level features and directly incorporate the essentials: e.g., the super-category of marine animals to detect sharks, or the super-category of plants to detect corn varieties. These task-oriented configurations are called 'experts'.

2 RELATED WORKS

Hierarchical networks (Yan et al., 2015) are gaining interest as Xue et al. (2022) showed their ability to improve model performance and efficiency. There are two key differences from other hierarchical networks: 1) Branches are not connected by fully connected (FC) layers or other operations in a

*Niels Hannemann, Bruno Heberle, and Lucas Schmidt contributed equally to this work as part of their bachelor thesis.



(a) Conventional (b) Hierarchical (NHL) (c) Efficient expert

Figure 1: Neural network topologies (blue: tensors, black: connection path). a conventional deep network from left to right processing (the dots represent five categories); b Our proposed nested structure; c Our innovative expert extraction technique.

later step, only by softmax! 2) Nesting is supported by a large language model. To the best of our knowledge, our approach of generating efficient experts by selecting high-level features is unique.

3 METHOD

Integration of the hierarchy in a classic CNN (c.f. Fig. 1a): The last conv block, the fifth in ResNet50 and ConvNeXtV2, is deleted and the FC is removed. The first branching point for the hierarchy is followed, the 'split point'. This is the first of three branching options. We construct the **N**est features at a **H**igh-**L**evel (**NHL**) hierarchy (c.f. 1b) by estimating the similarity between the categories using a quick semi-automatic approach with OpenAI's ChatGPT4 (Brown et al., 2020) for all ImageNet's 1k categories; a single example: 1) marine animals, 2) sharks, 3) hammerhead shark. Channels decrease with each new hierarchy level: 1) 128, 2) 64, and 3) 32; while maintaining the bottleneck structure. Each branch end has its own FC, which contributes a value per category to softmax-log-loss, so training proceeds as for many other classification tasks. Once trained our proposal can be utilized for any combination of the learned categories, here named as expert. ImageNet100 provides $\mathbf{2^{100} - 1}$ and ImageNet1k $\mathbf{2^{1000} - 1}$ possible experts (c.f. 1c), which do not require retraining.

4 EXPERIMENTS

To show our method's flexibility, we used a ResNet50 (He et al., 2016), trained from scratch, and a pre-trained ConvNeXtV2 (Woo et al., 2023). The ResNet50$_{\mathbf{NHL}}$ expert for 20 categories shows in Tab. 1 that despite similar GMACs and 41.3 % fewer parameters, it has a 2.8% top-1 acc. improvement over baselines. Although the NHL with 1000 categories has become a bit large (in parallel), its expert with 5 categories shows the same top-1 acc. as a deeper, pre-trained and fine-tuned ConvNeXtV2, despite 13.2% fewer GMACs and 51.4% fewer parameters. Such task-specific adjustments demonstrate the adaptability of our method over baselines that are not applicable (N/A) in this a way.

Table 1: Comparison of ResNet50 and ConvNeXtV2 with our NHL and 2 exemplary experts.

	ImageNet100		**ImageNet1k**	
Metric	**ResNet50**	**ResNet50$_{\mathrm{NHL}}$**	**ConvNeXtV2**	**ConvNeXtV2$_{\mathrm{NHL}}$**
Top-1 acc.	85.3%	**85.7%**	80.1%	**81.1%**
GMACs	**4.13**	5.86	**4.46**	14.55
GMACs of expert	N/A	4.22	N/A	**3.87**
GMACs reduction	0%	**-28%**	0%	**-73.4%**
Parameter	**23.5 / 23.7M**	25M	**27.8 / 28.6M**	119.9M
Parameter of expert	N/A	**13.8M**	N/A	**13.5M**
Parameter reduction	0%	**-44.5%**	0%	**-88.7%**
Train categories	20 / 100	100	5 / 1000	1000
Val categories	20	20	5	5
Top-1 acc.	88.5 / 90.6%	**93.4%**	**80.0** / 79.6%	**80.0%**

5 CONCLUSION

We introduce a novel principle for arbitrarily matching computational complexity to task complexity without compromising predictive performance. This method is especially promising for mobile computing, industrial, drone, robotics, and edge device applications, where very specific tasks are defined such that the relevant high-level features can be dynamically selected as needed, utilizing our method. Our experts have already achieved better top-1 accuracy with similar GMACs or the same top-1 accuracy with fewer GMACs, but we are confident that optimizing the base NHL model will unlock even more potential for low-resource computing.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. 2

Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. A survey on deep neural network pruning-taxonomy, comparison, analysis, and recommendations, 2023. 1

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016. doi: 10.1109/CVPR.2016.90. 2

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, June 2023. 1, 2

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html. 1

Mengqi Xue, Jie Song, Li Sun, and Mingli Song. Tree-like branching network for multi-class classification. In Pandian Vasant, Ivan Zelinka, and Gerhard-Wilhelm Weber (eds.), *Intelligent Computing & Optimization*, pp. 175–184, Cham, 2022. Springer International Publishing. ISBN 978-3-030-93247-3. 1

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2740–2748, 2015. doi: 10.1109/ICCV.2015.314. 1