Partial Least Squares: A Deep Space Odyssey

Artur Jordão Lima Correia[1] **and William Robson Schwartz**[1]

[1]Instituto de Ciências Exatas – Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – Brazil

{arturjordao,william}@dcc.ufmg.br

Abstract. *Modern visual pattern recognition models are based on deep convolutional networks. Such models are computationally expensive, hindering applicability on resource-constrained devices. To handle this problem, we propose three strategies. The first removes unimportant structures (neurons or layers) of convolutional networks, reducing their computational cost. The second inserts structures to design architectures automatically, enabling us to build high-performance networks. The third combines multiple layers of convolutional networks, enhancing data representation at negligible additional cost. These strategies are based on Partial Least Squares (PLS) which, despite promising results, is infeasible on large datasets due to memory constraints. To address this issue, we also propose a discriminative and low-complexity incremental PLS that learns a compact representation of the data using a single sample at a time, thus enabling applicability on large datasets.*

1. Introduction

Pattern recognition plays an important role in cognitive tasks such as natural language processing and image understanding. Modern pattern recognition methods have led to a series of breakthroughs, often surpassing human performance [Badia et al. 2020]. The reason for these remarkable achievements is the improvement in data representation (i.e., features), which allows discovering new abstractions and patterns from data.

In the context of visual pattern recognition, deep convolutional networks have been the focus of intense research due to their state-of-the-art effectiveness in learning discriminative representation. In particular, most efforts have been devoted to the development of architectures for convolutional networks, since large architectures are a major determinant factor for improving their predictive ability [Tan and Le 2019], as shown in Figure 1. In terms of performance, on the other hand, excessively large architectures are computationally expensive, hindering applicability on low-power and internet of things (IoT) devices. Moreover, such architectures are *data-hungry*, meaning that large datasets are needed to provide a better generalization performance [Kolesnikov et al. 2020], hence, the encouragement for large datasets has been growing.

A parallel line of research to obtain discriminative representations is to discover low-dimensional features through dimensionality reduction techniques. Such techniques are capable of yielding discriminative and compact representations from the original (high-dimensional) data [Li et al. 2019]. Recent works use dimensionality reduction collaboratively with convolutional networks and produce encouraging results [Suau et al. 2020]. Such a combination, however, is unsuitable for large datasets

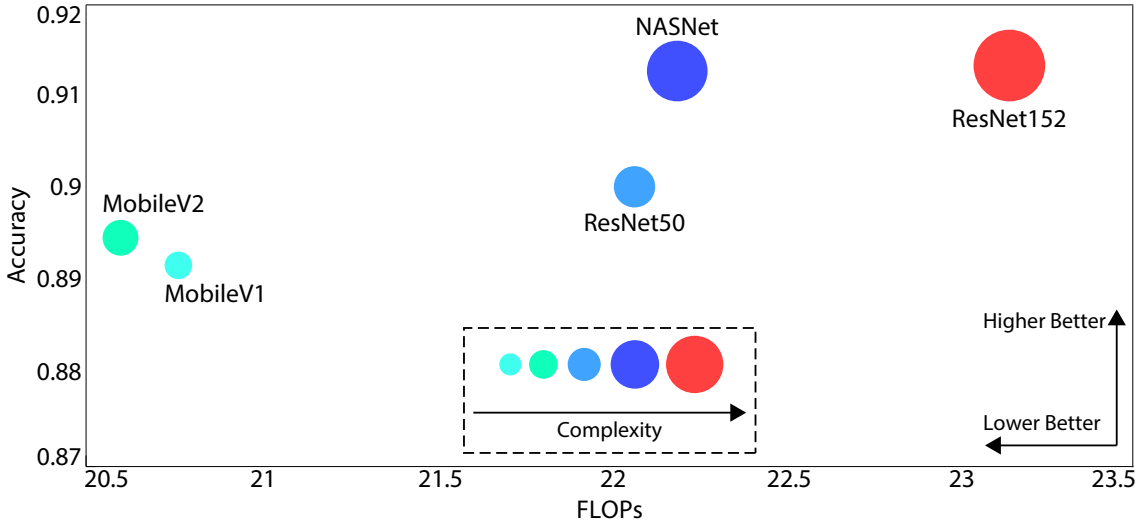

Figure 1. Comparison of convolutional networks in terms of predictive ability, computational cost, and complexity. Predictive ability is measured by accuracy. Computational cost is measured by Floating Point Operations (FLOPs). Complexity is measured taking into account the number of neurons (width) and layers (depth), and it is represented by the circle size (larger means more complex). The arrows indicate which direction (in both x and y axes) is better. It is evident that there is a strong relationship between predictive ability and network complexity (circle size), in which more complex networks are more accurate. In turn, network complexity incurs computational cost.

since traditional dimensionality reduction techniques require all the data to be in memory in advance, which is often impractical due to hardware limitations. Additionally, this requirement prevents us from employing dimensionality reduction on streaming applications, where the data are being generated continuously.

Regardless of the mechanism employed to recognize or improve pattern recognition, there is a trade-off between accuracy and complexity, in which more accurate models often incur higher complexity and computational cost, as illustrated in Figure 1. Thereby, discovering accurate and efficient strategies for pattern recognition, which include enhancing the existing ones, have been the focus of intense research.

Motivation. Modern visual pattern recognition models are predominantly based on convolutional networks since they are capable of learning effective representations from data [He et al. 2016]. According to previous works [Tan and Le 2019], large (deeper and wider) convolutional networks lead to better results. Figure 1 supports this claim, where larger networks (large circles) have superior predictive ability. In terms of performance, however, such networks suffer from heavy computation and memory overhead, incurring slow inference and hindering applicability on low-power and resource-constrained devices. Moreover, since modern networks demand massive computing infrastructure, the financial cost to deploy them can be prohibitive for academic researchers. For example, the estimated cost for training and deploying state-of-the-art networks can surpass hundreds of dollars per hour [Strubell et al. 2019]. Prior research on the climate impact of AI has raised another important issue regarding these networks, which is the quantity of

carbon emitted by them based on their energy consumption [Lacoste et al. 2019]. Surprisingly, convolutional networks have a large carbon footprint that can surpass one car in its lifetime [Strubell et al. 2019, Lacoste et al. 2019]. The simplest way to circumvent the problems mentioned is to evaluate different trade-offs between accuracy and network complexity, for example, by comparing the performance of ResNet50 (50 layers) with its deeper counterpart ResNet152 (152 layers), see Figure 1. Unfortunately, this process requires significant human engineering due to its trial-and-error essence. Instead, it is possible to transform or automatically design efficient convolutional networks by employing pruning or neural architecture search (NAS), respectively. The former removes unimportant structures (neurons or layers) from the network, reducing its complexity while preserving as much predictive ability as possible. The latter learns to design accurate and efficient architectures automatically. Both strategies, however, are not without their limitations. Existing criteria for identifying and removing structures from convolutional networks are ineffective since the accuracy of the original (unpruned) network is often degraded, as shown in Figure 2 (Left). Besides, many pruning approaches demand a high computational cost, mainly when applied to very deep networks [Luo and Wu 2020]. Regarding the neural architecture search, current strategies analyze a large set of possible candidate architectures and, hence, require vast computational resources and take many days to process even with modern Graphics Processing Units (GPUs) [Zoph et al. 2018]. Motivated by these issues, we propose simple, effective, and efficient mechanisms for eliminating structures of deep networks as well as discovering high-performance architectures automatically (i.e., without involving human engineering). More precisely, our pruning strategies achieve the best trade-offs between accuracy and computational cost compared to state-of-the-art methods, as illustrated in Figure 2 (Left). In the context of

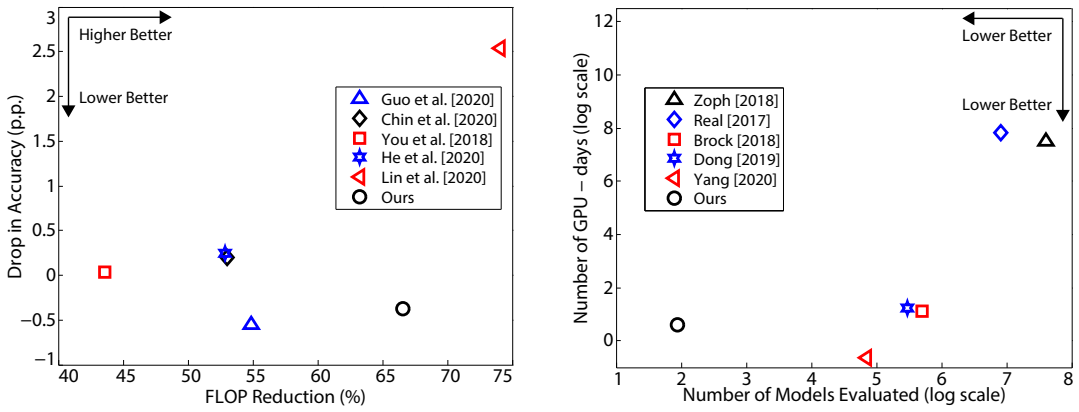

Figure 2. Left. Comparison of existing pruning methods. Compared to state-of-the-art pruning strategies, our pruning method always provides a better solution (i.e., it is a non-dominated solution) considering one of the performance metrics: accuracy drop (y-axis) or FLOP reduction (x-axis). In this figure, negative values in the y-axis denote improvement regarding the original, unpruned, network. Right. Comparison of existing neural architecture search (NAS) methods. Our NAS method discovers architectures by exploring one order of magnitude fewer models compared to other approaches. In addition, our method is the most resource-efficient as it designs architectures in a few hours on a single GPU. In both figures, the arrows indicate which direction is better.

NAS, our method discovers competitive and low-cost convolutional networks by exploring one order of magnitude fewer models compared to other approaches, thus designing architectures in a few hours on a single GPU, as shown in Figure 2 (Right).

Besides computational cost concerns, many efforts have been devoted to improve data representation of convolutional networks. In this context, previous works have demonstrated encouraging results combining features from different levels (layers) of the network. Such works have followed either multi-scale or HyperNet strategies. While the former redesigns network topology to encode features from shallow and deep layers [Yang et al. 2020], the latter preserves network topology, encouraging application on off-the-shelf networks [Sindagi and Patel 2019]. Despite the positive results, both strategies increase the computational burden significantly since they insert time-consuming operations at multiple levels of the network. To address this problem, we propose an efficient yet accurate approach to extract different levels of representation across multiple layers of deep networks, thus enhancing data representation at negligible additional cost.

A parallel line of research to improve data representation is to learn compact, but discriminative, representations through dimensionality reduction [Li et al. 2019]. In this context, Partial Least Squares (PLS) has presented remarkable results, mainly when compared to other methods such as Principal Component Analysis and Linear Discriminant Analysis [Sharma and Jacobs 2011]. The promising results of PLS are associated with its characteristics that include being discriminative and robust to *sample size problem* (when the number of samples is smaller than the number of features). Another attractive aspect of PLS is that it can operate as a feature selection method. However, PLS is unsuitable for large datasets (e.g., ImageNet) since all the data need to be available in advance and this could be impractical due to memory constraints. To handle this problem, many works have proposed incremental versions of traditional dimensionality reduction methods [Zeng and Li 2014], where the idea is to learn compact representations using a single sample at a time. Unfortunately, most incremental PLS fail to keep all its properties and present a high time complexity. To preserve the fundamental characteristics of PLS, we propose a discriminative and low-complexity incremental PLS. Among the advantages of this approach are the preservation of discriminative information, its computational efficiency, and the ability to operate as a feature selection technique.

Objectives. From a practical perspective, our goal is to promote mechanisms capable of reducing the financial cost, carbon emission and computational cost of convolutional networks (see Figure 3). More specifically, we pretend to provide strategies for (i) accelerating convolutional networks, (ii) discovering high-performance convolutional architectures automatically and (iii) efficiently improving data representation of convolutional networks. Additionally, we target to provide a memory-friendly version of PLS. From a theoretical perspective, our goal is to demonstrate the potential of PLS as a tool for determining the importance of structures composing a convolutional network. Besides, we intend to show that it is possible to preserve underlying properties of PLS in its incremental version through simple algebraic decomposition.

Contributions. The contributions of this thesis are simple, effective and efficient strategies for improving computational cost and predictive ability of convolutional networks. Specifically, we reduce more than half of computation, memory usage and inference time,

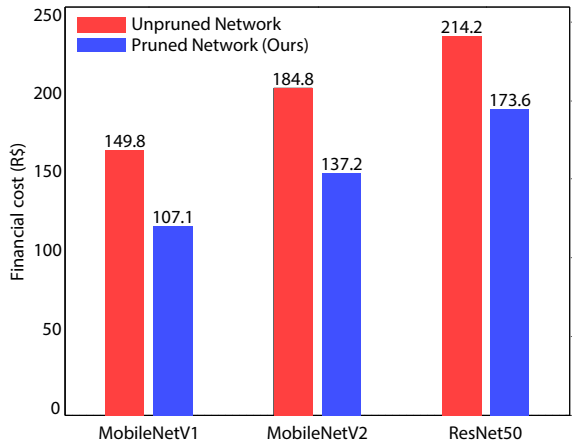

Figure 3. Financial cost (Brazilian real) and carbon emission for training different convolutional networks. Values above the bars indicate CO_2 in $kgCO_2eq$ (lower is better), which indicates the global warming potential of various greenhouse gases as a single number. Our strategies (blue bars) provide significantly more efficient convolutional networks.

which enables modern convolutional networks suitable to low-power systems (we refer the reader to Tables 5.13, 5.16, and Figures 5.10, 5.16 in the thesis for additional details). Furthermore, we decrease the financial cost of deploying convolutional networks, which is significant progress in making them more accessible to academic researchers, as shown in Figure 3. Regarding the climate impact of AI, our work enables that modern networks emit around 91% less CO_2. This result is an important step towards green AI. Last but not the least, we expand the applicability of a powerful dimensionality reduction technique, PLS, to large datasets and streaming applications. Particularly, all our contributions are beneficial for academics, researchers, and students with limited computational budgets. To promote reproducibility, we release the source code at: https://arturjordao.github.io/PLSDeepSpaceOdyssey/.

Publications. The results obtained during our research have been published in important conferences and journals on computer vision and pattern recognition:

1. Jordao, A., Yamada, F., and Schwartz, W. R. Deep Network Compression based on Partial Least Squares. Neurocomputing, 2020.
2. Jordao, A., Lie, M., and Schwartz, W. R. Discriminative Layer Pruning for Convolutional Neural Networks. Journal of Selected Topics in Signal Processing, 2020.
3. Jordao, A., Kloss, R. B., and Schwartz, W. R. Latent hypernet: Exploring the layers of Convolutional Neural Networks. International Joint Conference on Neural Networks (IJCNN), 2018.
4. Jordao, A., Kloss, R., Yamada, F., and Schwartz, W. R. Pruning Deep Neural Networks using Partial Least Squares. British Machine Vision Conference (BMVC) Workshops: Embedded AI for Real-Time Machine Vision, 2019.
5. Jordao, A., Yamada, F., Lie, M., and Schwartz, W. R. Stage-Wise Neural Architecture Search. International Conference on Pattern Recognition (ICPR), 2020.
6. Jordao, A., Lie, M., de Melo, V. H. C., and Schwartz, W. R. Covariance-free partial least squares: An Incremental Dimensionality Reduction Method. Winter Conference on Applications of Computer Vision (WACV), 2021.

Acknowledgments. The authors would like to thank Ricardo Kloss, Maiko Lie, Fernando Yamada, and Victor de Melo for their valuable contributions to this thesis. The authors would like to thank the Brazilian National Research Council – CNPq (Grants 438629/2018-3, 309953/2019-7 and 140082/2017-4), the Minas Gerais Research Foundation FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

References

Badia, A. P., Guoand, B. P. S. K. P. S. A. V. Z., and Blundell, C. (2020). Agent57: Outperforming the atari human benchmark. In *International Conference on International Conference on Machine Learning (ICML)*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision (ECCV)*.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. In *Neural Information Processing Systems (NeurIPS)*.

Li, Y., Yang, M., and Zhang, Z. (2019). A survey of multi-view representation learning. *Transactions on Knowledge and Data Engineering*, 31(10):1863–1883.

Luo, J.-H. and Wu, J. (2020). Neural network pruning with residual-connections and limited-data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sharma, A. and Jacobs, D. W. (2011). Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sindagi, V. and Patel, V. M. (2019). Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *International Conference on Computer Vision (ICCV)*.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Conference of the Association for Computational Linguistics*.

Suau, X., Zappella, L., and Apostoloff, N. (2020). Filter distillation for network compression. In *Winter Conference on Applications of Computer Vision (WACV)*.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*.

Yang, L., Han, Y., Chen, X., Song, S., Dai, J., and Huang, G. (2020). Resolution adaptive networks for efficient inference. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zeng, X. and Li, G. (2014). Incremental partial least squares analysis of big streaming data. *Pattern Recognition*, 47:3726–3735.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.