ERCIM fellowship Programme

Final scientific report

ERCIM Alain Bensoussan Fellowship Programme

ERCIM European Research Consortium for Informatics and Mathematics

Fellow	Dr. Shrutika Shankar Sawant
Host Organisation	Fraunhofer IIS, Erlangen, Germany
Scientific coordinator	Dr. Dr. Theresa Götz

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

1. Designed and developed novel filter pruning approaches for compressing over-parameterized models, which not only reduces storage space but also accelerates the inference of the Convolutional Neural Network model.

   Brief description about the activity:

   In recent years, deep convolutional neural networks (CNN) have evolved significantly in order to demonstrate remarkable performance in various computer vision tasks. However, the success of CNN comes at the cost of a vast amount of parameters and heavy computations hampering their deployment on mobile or embedded devices with limited computational resources. This problem has motivated the research community to investigate effective approaches that can reduce computational burden without compromising its performance. A prominent solution is to perform network compression without ominously degrading the models performance. Filter pruning has been recognized as a useful technique to compress and accelerate the CNNs, where weak or unimportant convolutional filters are eliminated. During the fellowship period, we have proposed different approaches for pruning the structures (i.e., filters) in deep CNNs without affecting their accuracy through learning only the important filters. Extensive experiments on several widely known CNN models for various applications (specifically, classification and segmentation) verify that our pruning approaches can efficiently compress CNN models with almost negligible or no loss of accuracy.
2. Wrote scientific articles in international peer-reviewed journals, as well as the highly technical replies to the questions raised by the reviewers.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Journals:

1. Shrutika S. Sawant, J. Bauer, F. X. Erick, Subodh Ingaleshwar, N. Holzer, A. Ramming, E. W. Lang & Th. Götz, 2022. "An optimal-score-based filter pruning for deep convolutional neural networks". Applied Intelligence, 52, pages 17557–17579 (2022). (Springer) https://doi.org/10.1007/s10489-022-03229-5. **Published.**

   Abstract: Convolutional Neural Networks (CNN) have achieved excellent performance in the processing of high-resolution images. Most of these networks contain many deep layers in quest of greater segmentation performance. However, over-sized CNN models result in overwhelming memory usage and large inference costs. Earlier studies have revealed that over-sized deep neural models tend to deal with abundant redundant filters that are very similar and provide tiny or no contribution in accelerating the inference of the model. Therefore, we have proposed a novel optimal-score-based filter pruning (OSFP) approach to prune redundant filters according to their relative similarity in feature space. OSFP not only speeds up learning in the network but also eradicates redundant filters leading to improvement in the segmentation performance. We empirically demonstrate on widely used segmentation network models (TernausNet, classical U-Net and VGG16 U-Net) and benchmark datasets (Inria Aerial Image Labeling Dataset and Aerial Imagery for Roof Segmentation (AIRS)) that computation costs (in terms of Float Point Operations

(FLOPs) and parameters) are reduced significantly, while maintaining or even improving accuracy.

2. Shrutika S. Sawant, Marco Wiedmann, Stephan Göb, N. Holzer, E. W. Lang & Th. Götz, 2022. “Compression of Deep Convolutional Neural Network Using Additional Importance-Weight-Based Filter Pruning Approach”. Applied Sciences, 12, 11184. (MDPI) https://doi.org/10.3390/app122111184 . **Published.**

   Abstract: The success of the convolutional neural network (CNN) comes with a tremendous growth of diverse CNN structures, making it hard to deploy on limited-resource platforms. These over-sized models contain a large amount of filters in the convolutional layers, which are responsible for almost 99% of the computation. The key question here arises: Do we really need all those filters? By removing entire filters, the computational cost can be significantly reduced. Hence, in this article, a filter pruning method, a process of discarding a subset of unimportant or weak filters from the original CNN model, is proposed, which alleviates the shortcomings of over-sized CNN architectures at the cost of storage space and time. The proposed filter pruning strategy is adopted to compress the model by assigning additional importance weights to convolutional filters. These additional importance weights help each filter learn its responsibility and contribute more efficiently. We adopted different initialization strategies to learn more about filters from different aspects and prune accordingly. Furthermore, unlike existing pruning approaches, the proposed method uses a predefined error tolerance level instead of the pruning rate. Extensive experiments on two widely used image segmentation datasets: Inria and AIRS, and two widely known CNN models for segmentation: TernausNet and standard U-Net, verify that our pruning approach can efficiently compress CNN models with almost negligible or no loss of accuracy. For instance, our approach could significantly reduce 85% of all floating point operations (FLOPs) from TernausNet on Inria with a negligible drop of 0.32% in validation accuracy. This compressed network is six-times smaller and almost seven-times faster (on a cluster of GPUs) than that of the original TernausNet, while the drop in the accuracy is less than 1%. Moreover, we reduced the FLOPs by 84.34% without significantly deteriorating the output performance on the AIRS dataset for TernausNet. The proposed pruning method effectively reduced the number of FLOPs and parameters of the CNN model, while almost retaining the original accuracy. The compact model can be deployed on any embedded device without any specialized hardware. We show that the performance of the pruned CNN model is very similar to that of the original unpruned CNN model. We also report numerous ablation studies to validate our approach.
3. Shrutika S. Sawant, F. X. Erick, Stephan Göb, N. Holzer, E. W. Lang & Th. Götz, 2022. “An Adaptive Binary Particle Swarm Optimization for Solving Multi-objective Convolutional Filter Pruning Problem”, **under review** in The Journal of Supercomputing.
4. Shrutika S. Sawant, Ashutosh Singh, Stephan Göb, N. Holzer, E. W. Lang & Th. Götz, 2022. “KlienesUNet: A Lightweight U-Net Based on Dilated Depth wise Separable Convolution for Semantic Segmentation”, **In process**.

Conferences:

1. Erick, F., Shrutika S. Sawant, Göb, S., Holzer, N., Lang, E. and Götz, T. (2022). A Simple and Effective Convolutional Filter Pruning based on Filter Dissimilarity Analysis. **Presented and published** In Proceedings of the 14th International Conference on Agents and Artificial

Intelligence - Volume 3: ICAART, ISBN 978-989-758-547-0; ISSN 2184-433X, pages 139-145. DOI: 10.5220/0010786400003116

Abstract: In this paper, a simple and effective filter pruning method is proposed to simplify the deep convolutional neural network (CNN) and accelerate learning. The proposed method selects the important filters and discards the unimportant ones based on filter dissimilarity analysis. The proposed method searches for filters with decent representative ability and less redundancy, discarding the others. The representative ability and redundancy contained in the filter is evaluated by its correlation with currently selected filters and left over unselected filters. Moreover, the proposed method uses an iterative procedure, so that less representative filters can be discarded evenly from the entire model. The experimental analysis confirmed that a simple filter pruning method can reduce floating point operations (FLOPs) of TernausNet by up to 89.65% on an INRIA Aerial Image Labeling dataset with an only marginal drop in the original accuracy. Furthermore, the proposed method shows promising results in comparison with other state-of-the-art methods.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

I. **Training programs:**

1. Business German for Beginners- Ongoing (Started on 21st September, 2021 Fraunhofer IIS- online mode)

2. Basic Qualification: Project management (Completed 30th - 31st of March 2021, Fraunhofer IIS)

3. Time management (Completed 22nd – 23rd of April 2021, Fraunhofer IIS)

II. **ERCIM FP Community event, 9th November, 2021 and 4th November, 2022.**

IV – RESEARCH EXCHANGE PROGRAMME (REP)

Research Exchange Programme (REP) is one of the important experience of the fellowship that provided me the opportunity to meet the expert in the respective research domain.

I have completed One week REP under the guidance of Dr. Christophe Giraud, Professor at Institut de Mathématiques d'Orsay in CELESTE Team of Inria at Université Paris-Saclay, Paris, France from 14.03.22 to 18.03.22.

Week started with interactive sessions with Dr. Christophe Giraud and fruitful discussions with his team. Exchange of thoughts with respect to the proposed objectives of the Research Exchange Program.

At the end of the REP week, I have given a presentation highlighting the scope of the collaborative work in the sparse Artificial Intelligence (AI) domain. I am positive about the solutions suggested by Dr. Christophe Giraud to the research problem. Overall, it was wonderful experience of conversing with Dr. Christophe Giraud.