Neural Architecture Optimization for Edge Devices

Neural Architecture Optimization for Edge Devices is a research project focused on developing techniques to optimize neural network architectures specifically for deployment on resource-constrained edge devices. Edge devices such as smartphones, IoT devices, or embedded systems often have limited computational power, memory, and energy resources. Therefore, designing lightweight and efficient neural networks becomes crucial to enable high-performance inference while minimizing resource usage.

This project involves investigating various methods to achieve this goal. Here are some key areas to explore:

1. Model Compression: Model compression techniques aim to reduce the size of neural network models without significant loss in performance. This can be achieved through techniques like weight pruning, quantization, and low-rank approximation. The research project could explore these techniques and their combinations to optimize the network's memory usage while preserving its functionality.
2. Architecture Search for Efficiency: Investigate methods for automatically searching for neural network architectures that are specifically tailored for edge devices. This could involve using techniques like reinforcement learning, genetic algorithms, or Bayesian optimization to explore the design space of architectures and identify ones that are both lightweight and efficient.
3. Knowledge Distillation: Knowledge distillation is a technique where a large, complex model (teacher model) is used to train a smaller, simplified model (student model) by transferring the knowledge from the teacher model to the student model. This approach can be utilized to create compact models that maintain high performance by leveraging the information learned by larger models.

References

[1] Besbes MD, Tabia H, Kessentini Y, Hamed BB. Progressive Learning With Anchoring Regularization For Vehicle Re-Identification. In2021 IEEE International Conference on Image Processing (ICIP) 2021 Sep 19 (pp. 1154-1158). IEEE.

[2] Mahmoudi MA, Chetouani A, Boufera F, Tabia H. Taylor series Kernelized layer for fine-grained recognition. In2021 IEEE International Conference on Image Processing (ICIP) 2021 Sep 19 (pp. 1914-1918). IEEE.

[3] Heuillet A, Tabia H, Arioui H, Youcef-Toumi K. D-DARTS: Distributed differentiable architecture search. arXiv preprint arXiv:2108.09306. 2021 Aug 20.

[4] Mahmoudi MA, Chetouani A, Boufera F, Tabia H. Deep kernelized network for fine-grained recognition. InNeural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III 28 2021 (pp. 100-111). Springer International Publishing.

[5] Mahmoudi MA, Chetouani A, Boufera F, Tabia H. Kernel-based convolution expansion for facial expression recognition. Pattern Recognition Letters. 2022 Aug 1;160:128-34.

[6] Heuillet A, Tabia H, Arioui H. NASiam: Efficient Representation Learning using Neural Architecture Search for Siamese Networks. arXiv preprint arXiv:2302.00059. 2023 Jan 31.

[7] Mahmoudi M, Chetouani A, Boufera F, Tabia H. Kernel function impact on convolutional neural networks. arXiv preprint arXiv:2302.10266. 2023 Feb 20.

[8] Mahmoudi MA, Boufera F, Chetouani A, Tabia H. Expanding Convolutional Neural Network Kernel for Facial Expression Recognition. InArtificial Intelligence: Theories and Applications: First International Conference, ICAITA 2022, Mascara, Algeria, November 7–8, 2022, Revised Selected Papers 2023 Mar 18 (pp. 3-17). Cham: Springer Nature Switzerland.

[9] Heuillet A, Nasser A, Arioui H, Tabia H. Efficient Automation of Neural Network Design: A Survey on Differentiable Neural Architecture Search. arXiv preprint arXiv:2304.05405. 2023 Apr 11.

[10] Mahmoudi MA, Chetouani A, Boufera F, Tabia H. Learnable pooling weights for facial expression recognition. Pattern Recognition Letters. 2020 Oct 1;138:644-50.