Scopus

Documents

Latif, S.A.[a] , Sidek, K.A.[a] , Bakar, E.A.[b] , Hashim, A.H.A.[a]

Online Multimodal Compression using Pruning and Knowledge Distillation for Iris Recognition
(2024) *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 37 (2), pp. 68-81.

DOI: 10.37934/araset.37.2.6881

[a] Department of Electrical and Computer Engineering, Kuliyyah of Engineering, International Islamic University Malaysia, Selangor, Malaysia
[b] Department of Electrical Engineering, Sultan Azlan Shah Polytechnic, Perak, Malaysia

Abstract

Deep learning models have advanced to the forefront of image recognition tasks, resulting in high-performing but enormous neural networks with millions to billions of parameters. Yet, deploying these models in production systems imposes considerable memory limits. Hence, the research community is increasingly aware of the need for compression strategies that can reduce the number of model parameters and their resource requirement. Current compression techniques for deep learning models have limitations in efficiency and effectiveness, indicating that more research is required to develop more efficient and practical techniques capable of balancing the trade-offs between compression rate, computational cost, and accuracy. This study proposed a multimodal method by combining multimodal Pruning and Knowledge Distillation techniques for compressing the iris recognition model, which is the size constraint for many image recognition models. To maintain accuracy while shrinking the model's size, the models are trained, compressed, and further retrained in the downstream job. The analysis includes both fully connected and convolutional layers. Experimentally, the findings show that the proposed technique can achieve 91% accuracy, the same as the existing or original model. Besides that, the model compression can reduce the size of the model almost six times, from 529MB to 90MB, which is a significantly reduced rate. The primary outcome of this study is developing a CNN lightweight model for iris recognition technology that can be used on mobile devices and is resource constrained. © 2024, Semarak Ilmu Publishing. All rights reserved.

Author Keywords
iris recognition; knowledge distillation; Model compression; pruning

References

- Cheng, Yu, Wang, Duo, Zhou, Pan, Zhang, Tao
  (2017) *A survey of model compression and acceleration for deep neural networks*, arXiv preprint arXiv:1710.09282

- Chen, Xizi, Zhu, Jingyang, Jiang, Jingbo, Tsui, Chi-Ying
  Tight Compression: Compressing CNN Through Fine-Grained Pruning and Weight Permutation for Efficient Implementation
  (2022) *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42 (2), pp. 644-657.

- Chen, Xizi, Zhu, Jingyang, Jiang, Jingbo, Tsui, Chi-Ying
  Tight compression: compressing CNN model tightly through unstructured pruning and simulated annealing based permutation
  (2020) *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1-6.
  IEEE

- Dupuis, Etienne, Novo, David, O'Connor, Ian, Bosio, Alberto
  Sensitivity analysis and compression opportunities in dnns using weight sharing
  (2020) *2020 23rd International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS)*, pp. 1-6.
  IEEE

- Oguntola, Ini, Olubeko, Subby, Sweeney, Christopher
  Slimnets: An exploration of deep model compression and acceleration
  (2018) *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pp. 1-6.
  IEEE

- Hu, Qingqiao, Yin, Siyang, Ni, Huiyang, Huang, Yisiyuan
  An end to end deep neural network for iris recognition
  (2020) *Procedia Computer Science*, 174, pp. 505-517.

- Nguyen, Kien, Fookes, Clinton, Ross, Arun, Sridharan, Sridha
  Iris recognition with off-the-shelf CNN features: A deep learning perspective
  (2017) *IEEE Access*, 6, pp. 18848-18855.

- Jamil, Amirah Hanani, Yakub, Fitri, Azizan, Azizul, Roslan, Shairatul Akma, Zaki, Sheikh Ahmad, Ahmad, Syafiq Asyraff
  A Review on Deep Learning Application for Detection of Archaeological Structures
  (2022) *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 26 (1), pp. 7-14.

- Jalilian, Ehsaneddin, Hofbauer, Heinz, Uhl, Andreas
  Iris Image Compression Using Deep Convolutional Neural Networks
  (2022) *Sensors*, 22 (7), p. 2698.

- Vyas, Ritesh
  Towards adept hand-crafted features for ocular biometrics
  (2020) *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1-6.
  IEEE

- Zhang, Yabo, Ding, Wenrui, Liu, Chunlei
  Summary of convolutional neural network compression technology
  (2019) *2019 IEEE International Conference on Unmanned Systems (ICUS)*, pp. 480-483.
  IEEE

- Alqahtani, Ali, Xie, Xianghua, Jones, Mark W.
  Literature review of deep network compression
  (2021) *Informatics*, 8 (4), p. 77.
  MDPI

- Fernandes, Francisco E., Yen, Gary G.
  Pruning deep convolutional neural networks architectures with evolution strategy
  (2021) *Information Sciences*, 552, pp. 29-47.

- Tung, Frederick, Mori, Greg
  Deep neural network compression by in-parallel pruning-quantization
  (2018) *IEEE transactions on pattern analysis and machine intelligence*, 42 (3), pp. 568-579.

- Qian, Liuchen, Fu, Yuzhuo, Liu, Ting
  An Efficient Model Compression Method for CNN Based Object Detection
  (2018) *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 766-769.
  IEEE

- Marinó, Giosué Cataldo, Petrini, Alessandro, Malchiodi, Dario, Frasca, Marco
  Deep neural networks compression: A comparative survey and choice recommendations
  (2023) *Neurocomputing*, 520, pp. 152-170.

- Gheorghe, Ștefan, Ivanovici, Mihai
  Model-based weight quantization for convolutional neural network compression
  (2021) *2021 16th International Conference on Engineering of Modern Electric Systems (EMES)*, pp. 1-4.
  IEEE

- Yang, Hong, Zhang, Ya-sheng, Yin, Can-bin, Ding, Wen-zhe
  Ultra-lightweight CNN design based on neural architecture search and knowledge distillation: A novel method to build the automatic recognition model of space target ISAR images
  (2022) *Defence Technology*, 18 (6), pp. 1073-1095.

- Cai, Gaoyuan, Li, Juhu, Liu, Xuanxin, Chen, Zhibo, Zhang, Haiyan
  Learning and Compressing: Low-Rank Matrix Factorization for Deep Neural Network Compression
  (2023) *Applied Sciences*, 13 (4), p. 2704.

- Fang, Beihua, Lu, Yuanfu, Zhou, Zhisheng, Li, Zhihui, Yan, Yuwen, Yang, Linfeng, Jiao, Guohua, Li, Guangyuan
  Classification of genetically identical left and right irises using a convolutional neural network
  (2019) *Electronics*, 8 (10), p. 1109.

- Alaslni, Maram G., Elrefaei, Lamiaa A.
  Transfer learning with convolutional neural networks for iris recognition
  (2019) *Int. J. Artif. Intell. Appl*, 10 (5), pp. 47-64.

- Singh, Arun, Pandey, Akansha, Rakhra, Manik, Singh, Dalwinder, Singh, Gurasis, Dahiya, Omdev
  An Iris Recognition System Using CNN & VGG16 Technique
  (2022) *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 1-6.
  IEEE

- Almutiry, Omar
  Efficient iris segmentation algorithm using deep learning techniques
  (2022) *Journal of Electronic Imaging*, 31 (4), pp. 041202-041202.

- Alrifaee, Mustafa M.
  (2020) *A short survey of iris images databases*,
  SSRN 3616735

- Proença, Hugo, Alexandre, Luís A.
  UBIRIS: A noisy iris image database
  (2005) *Image Analysis and Processing– ICIAP 2005: 13th International Conference*, pp. 970-977.
  Cagliari, Italy, September 6-8, Proceedings 13, Springer Berlin Heidelberg, 2005

- Danlami, Muktar, Jamel, Sapiee, Ramli, Sofia Najwa, Deris, Mustafa Mat
  A framework for iris partial recognition based on Legendre wavelet filter
  (2019) *International Journal of Advanced Computer Science and Applications*, 10 (5).

- Klosterman, S.
  Overfitting, underfitting, and the bias-variance tradeoff
  (2019) *Towards Data Science*,

- Luo, Jian-Hao, Wu, Jianxin, Lin, Weiyao
  Thinet: A filter level pruning method for deep neural network compression
  (2017) *Proceedings of the IEEE international conference on computer vision*, pp. 5058-5066.

Correspondence Address
Sidek K.A.; Department of Electrical and Computer Engineering, Selangor, Malaysia; email: azami@iium.edu.my

Publisher: Semarak Ilmu Publishing

ISSN: 24621943
Language of Original Document: English
Abbreviated Source Title: J. Adv. Res. Appl. Sci. Eng. Technol.

2-s2.0-85183643439
Document Type: Article
Publication Stage: Final
Source: Scopus