Machine Vision in Resource-Constrained Embedded Systems

Timothy J Wiegman
Purdue University
Agricultural and Biological Engineering
West Lafayette, IN, USA
wiegman@purdue.edu

ABSTRACT

Modern machine learning now enables relatively advanced computer vision technology to be deployed on simple, low-resource computing hardware. This allows proliferation of an important type of artificial intelligence to many applications where cost or distribution would have previously been prohibitive. Three recent papers are discussed, showing how such technology is used and developed for embedded platforms. Though this is exciting, it is expected that even more progress is forthcoming both as machine vision software further matures and as advances in computer technology bring more powerful computation to wider distribution and lower cost.

ACM Reference Format:
Timothy J Wiegman. 2024. Machine Vision in Resource-Constrained Embedded Systems. In . ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Computer vision is an important application of artificial intelligence, a critical technology of the information age. The possible uses for this technology are extremely numerous, from automation to zoology and everything in between ([Szeliski 2022code]; [Smith, Smith, and Hansen 2021code]; [Manoharan 2020code]). However, many of the most popular implementations require powerful computers or other (expensive) specialized equipment, as they are focused on doing the most challenging tasks with the most cutting-edge methods. This means those are unsuited to distributed deployments, real-time requirements, or otherwise scaling beyond niche usage ([Bayoudh et al. 2022code]).

However, with the advent of modern machine learning methods, some simplified computer vision models can be run on inexpensive, widespread electronic devices using little more than basic microcontrollers ([Immonen and Hämäläinen 2022code]). Deploying computer vision technology in this fashion makes it more accessible to the general public, and it is more secure, private, and reliable than outsourcing computation to centralized cloud servers.

Purdue ECE568, Spring 2024, West Lafayette, IN, USA

2 REVIEW

2.1 Prescreening Oral Cancers with TensorFlow Lite For Microcontrollers

The first paper to review in detail is Shamim ([2022code]). This paper describes a medical application for TensorFlow Lite For Microcontrollers, a "TinyML" ([Immonen and Hämäläinen 2022co software that compresses traditional AI models to run on extremely basic ARM devices. The authors re-trained an existing model, originally designed for generic computer vision tasks on mobile devices, for their domain-specific goal of detecting and classifying tongue lesions that can develop into oral cancers. Once fine-tuned in this way, the model was quantized and compressed from 32-bit float to 8-bit integer precision. This converted model, despite being significantly simpler and requiring far less resources (63% lower peak memory and 80% smaller executable), still achieved nearly equivalent accuracy (within 1-2%) and latency (within 0.01 ms). All computation was performed on an ST Microelectronics STM32H743II, a relatively inexpensive and efficient 32-bit microprocessor based on ARM Cortex-M7.

2.2 Embedding Crosswalk Detection into a Wearable Device

The second paper to review in detail is Silva et al. ([2020code]). This paper describes the development of a wearable crosswalk detection device, optimized for minimal memory and power consumption. The first portion of the study shows several experiments used to quantify how different choices in the computer vision pipeline—such as the input resolution, color depth, and even the architecture of the machine learning algorithm—affect the memory footprint of the application. The latter portion of the study explained how the application utilized the resources on their embedded platform: the Texas Instruments TM4C123GH6PM, a very inexpensive and efficient ARM Cortex-M4 based microcontroller. The system consumed far less than a watt, and required barely two dozen kilobytes of RAM and flash, and was still able to achieve nearly 90% accuracy with respectable sub-second latency.

2.3 Automating Design of Small-Yet-Powerful Neural Networks

The third and final paper to review in detail is Liberis, Dudziak, and Lane ([2021code]). This paper describes a software that can automatically design the architecture for a neural network with the goal of fitting within extremely tight resource constraints, such as those of low-power microcontrollers with slow processors and very little memory

available. This is a technology that can enable other engineers to more easily develop AI-powered applications for inexpensive and distributed devices such as Internet-of-Things (IoT) or wearable devices. The authors report that their system can improve model accuracies by a few percent, reduce memory usage multiple-fold, or reduce the number of operations (a proxy for latency) by up to two-fold.

2.4 Comparing the Papers

The first two papers are the most similar. Both describe the development of a prototype embedded device to solve a specific problem. The first was slightly simpler, as it used a higher level computer vision system and merely compressed it for use on a low-power embedded computer. However, the second paper targeted a significantly lower-power device, so the authors of that study explained a much more in-depth process for carefully optimizing their application for their extreme resource constraints. The third and final paper was somewhat abstracted from the others, as it focused on explaining a technology one step higher in the stack. None of the three papers were directly related or cited each other.

3 DISCUSSION AND CONCLUSION

I think the types of technologies developed in this field are excellent examples of practical artificial intelligence. While high-power applications like ChatGPT are glossy and visible and exciting, they are unlikely to permeate everyday life like embedded systems do. The ability to integrate cutting-edge computer technologies into low-cost, simple devices will allow them to spread further and faster than expensive and powerful ones. They can solve practical problems, like the second paper demonstrated (in that case, it could be an assistive technology for the visually impaired), while there is also the possibility that they will solve problems we cannot even yet foresee.

The biggest challenges to further rollout of these kinds of technology are likely going to be the difficulty of developing for such tightly resource-constrained hardware, as shown in the in-depth optimization process explained in Silva et al. ([2020code]). However, those roadblocks will be eased as better toolchains are built (as in Liberis, Dudziak, and Lane ([2021code])) and as powerful computers become cheaper and more widespread purely due to hardware development.

REFERENCES

Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2022. "A Survey on Deep Multimodal Learning for Computer Vision: Advances, Trends, Applications, and Datasets." *The Visual Computer* 38 (8): 2939–70. https://doi.org/10.1007/s00371-021-02166-7.

Immonen, Riku, and Timo Hämäläinen. 2022. "Tiny Machine Learning for Resource-Constrained Microcontrollers." *Journal of Sensors* 2022 (November): e7437023. https://doi.org/10.1155/2022/7437023.

Liberis, Edgar, Łukasz Dudziak, and Nicholas D. Lane. 2021. "Micro-NAS: Constrained Neural Architecture Search for Microcontrollers." In *Proceedings of the 1st Workshop on Machine Learning and Systems*, 70–79. EuroMLSys '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3437984.3458836.

Manoharan, Dr Samuel. 2020. "Embedded Imaging System Based Behavior Analysis of Dairy Cow." *Journal of Electronics and Informatics* 2 (2): 148–54. https://irojournals.com/iroei/article/view/2/2/6.

Shamim, Mohammed Zubair M. 2022. "Hardware Deployable Edge-AI Solution for Prescreening of Oral Tongue Lesions Using TinyML on Embedded Devices." *IEEE Embedded Systems Letters* 14 (4): 183–86. https://doi.org/10.1109/LES.2022.3160281.

Silva, Elias T., Fausto Sampaio, Lucas C. da Silva, David S. Medeiros, and Gustavo P. Correia. 2020. "A Method for Embedding a Computer Vision Application into a Wearable Device." *Microprocessors and Microsystems* 76 (July): 103086. https://doi.org/10.1016/j.micpro.2020.103086.

Smith, Melvyn L., Lyndon N. Smith, and Mark F. Hansen. 2021. "The Quiet Revolution in Machine Vision - a State-of-the-Art Survey Paper, Including Historical Review, Perspectives, and Future Directions." *Computers in Industry* 130 (September): 103472. https://doi.org/10.1016/j.compind.2021.103472.

Szeliski, Richard. 2022. *Computer Vision: Algorithms and Applications*. Springer Nature.