Qualcomm Researchers Pioneer On-Device Learning

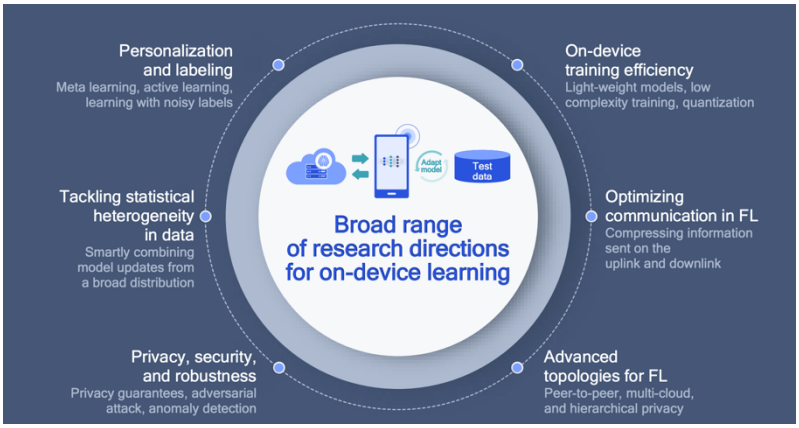
TRAINING ON MOBILE AND EDGE DEVICES COULD IMPROVE THE USER EXPERIENCE WHILE PRESERVING PRIVACY.

We've all had those frustrating experiences with our mobile phones when the voice assistant seems to possess artificial stupidity instead of artificial intelligence. The real problem here is that these assistants require cloud connectivity (increasing latency) and do not continuously learn from interactions with you over time. Qualcomm AI Research shows how to change that by adding on-device learning capabilities to mobile and edge devices, enabling personalization while preserving privacy.

THE CHALLENGE AND OPPORTUNITY OF ON-DEVICE LEARNING

Current devices are focused on inference processing rather than on-device learning, which is a much more difficult computational challenge in small, low-power devices. Moreover, existing techniques for training AI models not only use massive computational jobs, requiring trillions of floating-point calculations that can take days or weeks to complete on multiple racks of high-performance accelerators, but they also require large, centrally collected data sets. To accomplish this on a mobile phone will require entirely new approaches that consume less data, dramatically streamline the process, and extend to local private user data. Qualcomm AI Research has some promising ideas to address this and help achieve learning at the edge.

The technical problems are challenging but not insurmountable. We need to be able to run smaller models that adapt to the target data while preserving accuracy and privacy. Training with far less labeled data could better align with edge device memory, power, and compute capacity, and often personalization tasks need to fine-tune to specific target user examples. Furthermore, federated learning could be harnessed to pool edge resources for training across multiple devices. Computationally, we could reduce the numerical precision (quantization) required in training backprop to reduce compute and memory requirements, but only if mechanisms can be developed to maintain model accuracy. Qualcomm AI Research is pursuing several of these areas and more. See more details of the research directions in [this presentation by Qualcomm](#).



FEW-SHOT LEARNING

Learning from limited labeled data will be crucial to achieving the required accuracy when training on edge devices. The algorithmic approaches for efficient on-device learning can hinge on adapting models to deal with fewer labeled data samples, such as in few-shot learning or using unlabeled data in unsupervised learning in certain data domains. One can also target models with user-labeled data with limited sample sizes. Getting users to help label samples can also improve accuracy by

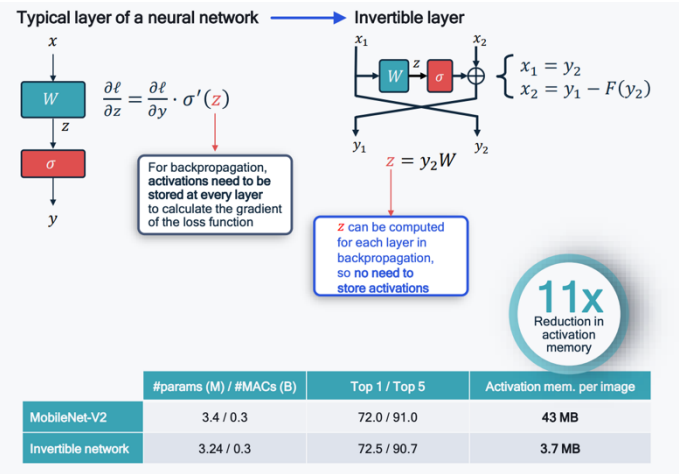
using cleaner data than is usually available from the environment. An excellent example of few-shot learning is improving keyword spotting (KWS). By adapting the model using data collected from user enrollment, one can personalize the model and significantly improve results.

Federated Learning (FL)

Much hope for on-device training has been pinned on a hybrid approach, pooling edge devices to aggregate model perturbations in the cloud and then distributing updated models back to the devices. Each device must adapt the model based on local data, then aggregate updates across multiple users globally for improved model performance across a broader range of observed data. One idea that also shows promise is to create embeddings (transforming a high-dimensional space into a lower-dimensional space) that are codewords of error-correcting codes to avoid the need to share individual embeddings over the network. This approach, called FedUV, has been demonstrated to deliver results comparable to more common techniques that share embeddings. Another challenge of FL is having adequate communication bandwidth and latency. While 5G certainly helps, sufficient bandwidth depends on compression/decompression algorithms, with good on-chip support.

Backpropagation Optimization

Using reduced precision integers and operators, quantization is becoming more commonplace in inference to avoid expensive floating-point calculations and reduce memory bandwidth requirements. Qualcomm AI Research is investigating two areas to reduce complexity without reducing model accuracy. One promising technique is "In-Hindsight Range Estimation," which predicts the level of precision required to maintain model accuracy when computing the gradients. The algorithm extracts statistics from the current tensor and applies that knowledge to calculate the quantization parameters for the next iteration. This lowers complexity and data movement by as much as 79%.



	#params (M) / #MACs (B)	Top 1 / Top 5	Activation mem. per image
MobileNet-V2	3.4 / 0.3	72.0 / 91.0	43 MB
Invertible network	3.24 / 0.3	72.5 / 90.7	3.7 MB

Another approach is to reduce memory requirements. Instead of storing every layer's activations to calculate the gradient of the loss function, Qualcomm AI Research is investigating using "invertible layers" to reduce memory for training significantly. For example, MobileNet-V2 training was accomplished with an 11x reduction in activation memory, from 43MB to 3.7 MB, without a meaningful reduction in prediction accuracy or change to the number of inference FLOPS or model parameters.

Conclusions

On-device learning on a mobile or edge device could allow developers to clear hurdles preventing some AI applications from achieving their full potential. But training on battery-operated power-efficient devices is tough to accomplish. Qualcomm AI Research believes it has meaningful solutions in flight, using not one silver bullet but a range of techniques that can be applied in isolation or harmony. This holistic approach is already showing promise in the lab and could revolutionize edge AI applications in the not-so-distant future.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR: Karl Freund, Founder, and Principal Analyst at Cambrian-AI Research

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be mentioned in the context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with QTI funding and support. Although the paper may utilize publicly available material from various vendors, including QTI, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The views expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.