EDGE IMPULSE ACCELERATES MLOPS WITH EON TUNER

Abstract

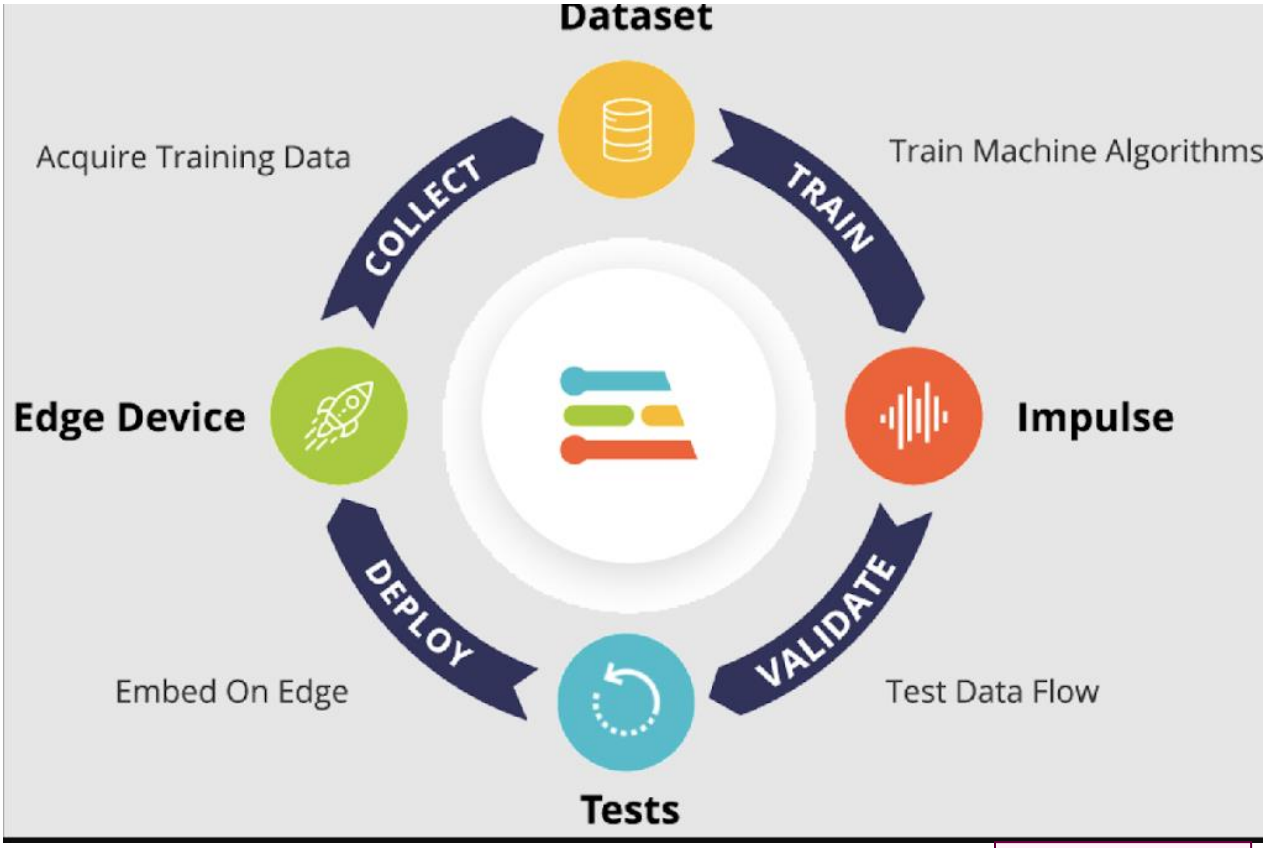
Edge Impulse's EON platform is a new tool for system developers.

SPONSORED BY Edge Impulse

Introduction

Building machine learning (ML) into embedded devices is more than a one-time event. Integrating machine learning is a process requiring iterative improvements overtime. In order to support a continuous process of training deployment feedback and retraining, it requires that machine learning operations (MLOps) be an integral part of any toolchain. It is therefore critical that even when developing ML for low power embedded devices, MLOps best practices are being followed.

This white paper explores how Edge Impulse will accelerate MLOps using the new "EON" capabilities. Edge Impulse is an end-to-end embedded ML platform with MLOps built in from the start, that allows developers to scale up from low power microcontrollers to high performance applications processors. What's important for intelligent edge designs is the ability to manage the ML solution for the entire lifecycle of the product from design to deployment. Especially considering how in an embedded environment, where debugging can be hard, it is important to support developers in catching errors and bugs earlier in the process and allow them to iterate more efficiently.



Source: Edge Impulse

Figure 1. MLOps production cycle.

MLOps includes all the capabilities that embedded engineers, ML developers, product teams, and IT operations need to deploy, manage, govern, and secure machine learning and other

probabilistic models running on embedded devices, in production. It combines the practice of AI/ML development with the principles of DevOps to define an ML lifecycle that exists in tandem to the software development lifecycle (SDLC). Its purpose is to support the continuous integration, development, and delivery of AI/ML models into production at scale and over time.

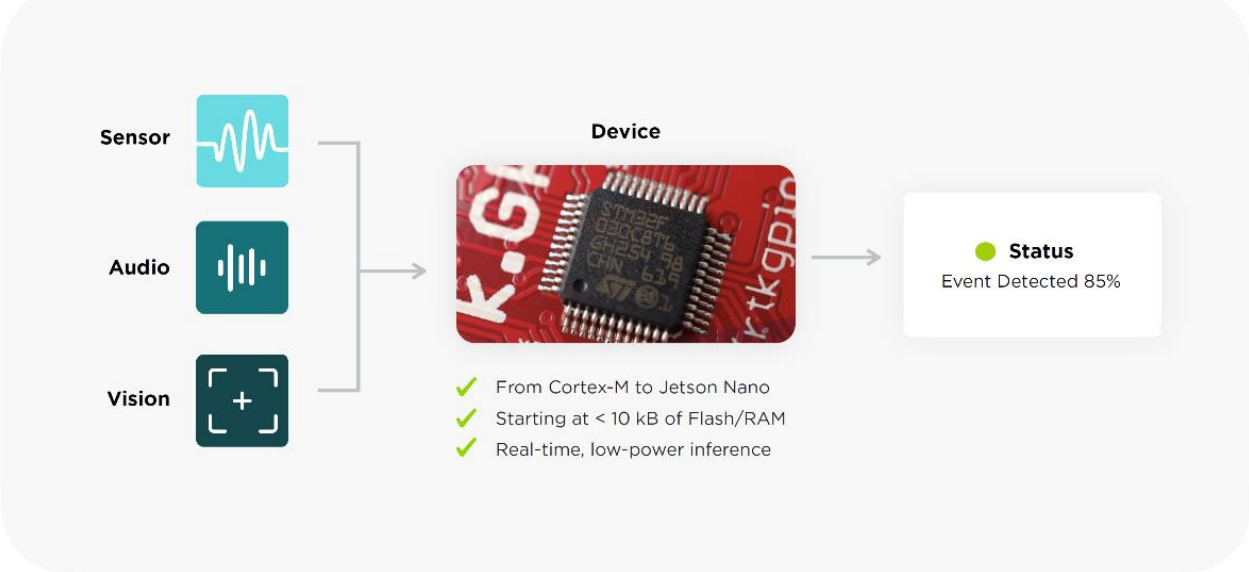
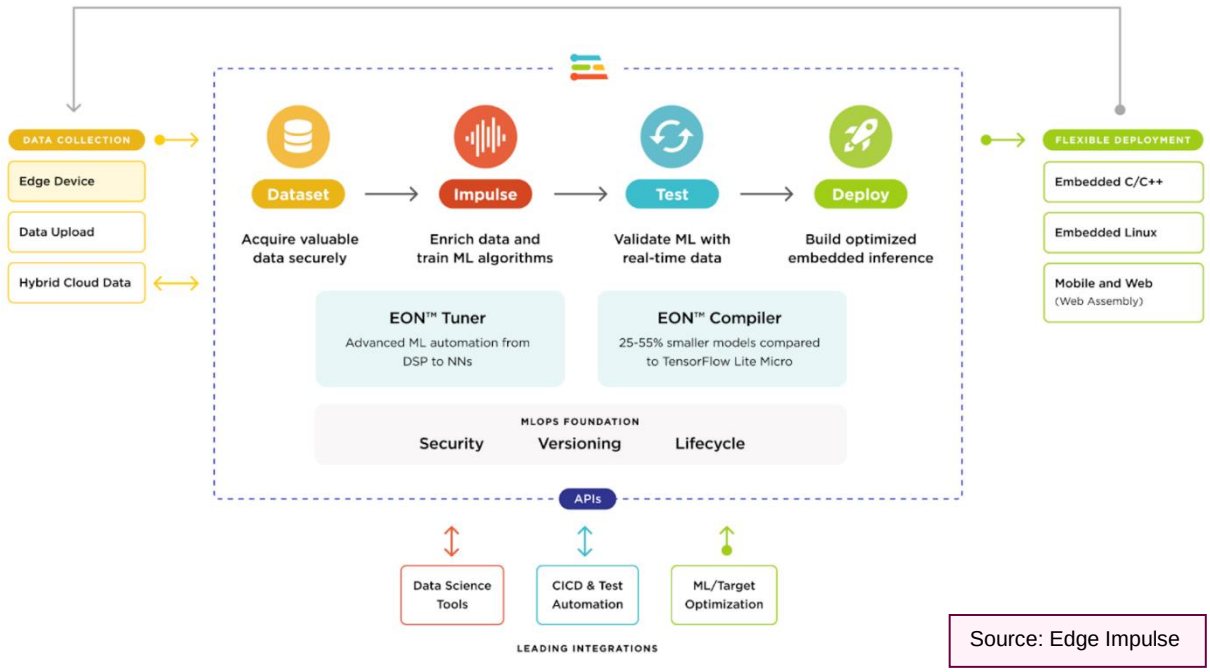


Figure 2. Adding Machine Learning to embedded edge devices make them smarter Source: Edge Impulse

Signal Processing in Edge Devices

The advantage that Edge Impulse has is that it allows developers to design, pick and build more compressed ML models that are easier and cheaper to deploy. This compressed model can provide real benefit over the entire life-cycle by making it cheaper and easier to make continual improvements.

In addition, Edge Impulse has integrated a variety of state of the art digital signal processing (DSP) front end algorithms to improve the raw input data to reduce the size of the ML model needed, simplifying the work of engineering teams. The sensor pipeline is a flow-through architecture with a combination of different tasks. The DSP front end functions include feature extraction, edge detection, signal smoothing, and de-noising. For example, feature extraction algorithms are helpful in making object detection or sound recognition more efficient while also increasing the accuracy.



Source: Edge Impulse

Figure 3. MLOps can make the continual model improvements and scale up to more complex platforms

Accelerating ML Design with AutoML

AutoML research has focused on how to get designers the best results without a lot of manual trial and error. Important criteria include processing time (latency), memory use, and storage requirements. These are especially important when running these ML applications on constrained devices such as microcontrollers. In a voice recognition example application, design factors include windows length, processing steps, down-sampling audio, in addition to the many hyperparameters associated with training a machine learning model. With so many options, a tool to help automate those decisions can be very valuable. One such tool is Edge Impulse's Edge Optimized Neural (EON) Tuner. [https://www.edgeimpulse.com/blog/introducing-eon]

While the EON Tuner is not completely automatic - i.e. with no user interaction - it is more like "assisted ML." And it looks at all the processing elements, not just the ML part, to evaluate the optimal solution. As such, the tool helps the programmer make an informed decision. But ultimately the programmer chooses the mix of DSP and ML elements, based on data. The EON tuner and compiler convert interpretive ML into compiled ML for faster execution and lower memory usage.

EON Tuner was specifically designed for sensor data applications like vibration, sounds, image, and short video, as those are the most common application for embedded designs. The EON

Tuner, given a set of constraints, can even determine whether it's more efficient to focus more on the feature extraction or on the ML stage and find the best compromise.

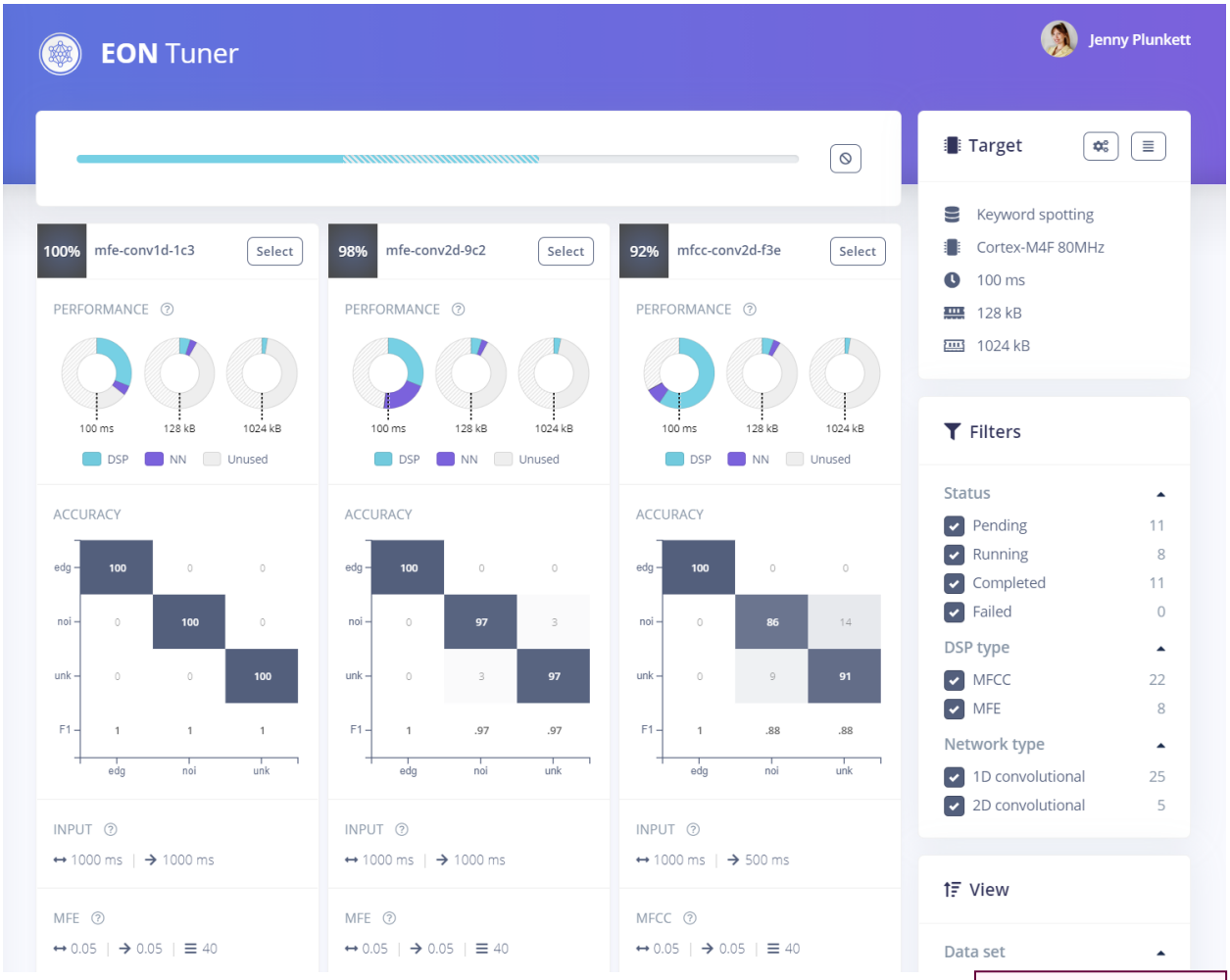


Figure 4. EON runs multiple scenarios to solve for the best fit

Source: Edge Impulse

For developers using one of many qualified target CPUs, EON Tuner is aware of its specific processor, memory, and storage configurations and can base its recommendations on this information when searching for the most effective ML pipeline. Once the target platform is selected, the designer can choose the total inference time limits. The EON Tuner will run through relevant DSP and ML solutions and provide 30 different ML pipeline outputs.

The EON Tuner will select the proper DSP pre-processor for the type of signal – continuous, or impulse. Continuous sampling requires windowing to limit the processing time. The ML section will have different convolutional neural nets including 1D or 2D convolutions for audio applications. The tool will suggest a neural net model based on the application. It can also start with a default fully-connected NN and then modify the model after training on supplied data.

The results are nominally ranked by accuracy achieved on the validation data set, with each offering variable DSP and NN times, memory required and storage requirements. One can then

decide to look at the performance on the train or test data sets to pick the best solution for their use case. The output also includes a spectrogram and the user can drill down into the FFT processor and NN layers. The user can even connect a live device to the cloud processing model to stream real life data on the model. In this case the cloud model acts as a digital twin. Using this capability, allows continual model testing and improvements to deploy in the field.

Edge Impulse platform gives the designer flexibility and produces an efficient model. The EON Compiler has produced solutions that ran the neural networks in 25-55% less RAM, and up to 35% less flash, while retaining the same accuracy, compared to TensorFlow Lite for Microcontrollers

Conclusion

To get ML into the hands of a typical embedded engineer requires a new set of tools. Several startups, including Edge Impulse, are trying to make the process of building ML into embedded systems accessible to a much wider range of developers with end-to-end tools with “low-code” to advanced capabilities all in one tool.

With the Edge Impulse EON Tuner, programmers can more quickly make design decisions and tradeoffs and fit more complex functions into smaller devices. The EON Tuner tool is just a small part of the full MLOps solution offered by Edge Impulse to speed your product development.