BrainChip Sees AI Gold in Sequential Data Analysis at the Edge

Karl Freund
Cambrian-AI Research, LLC
August 22, 2023

Unlike image processing or large language models, few AI startups are focused on sequential data processing, which includes video processing and time-series analysis. BrainChip is just fine with that.

With all the buzz around LLM generative AI, it is understandable that other forms of AI seem to have vaporized in the ChatGPT mist. One often overlooked area is analyzing time-series data, such as streaming stock quotes and video processing. BrainChip has singled out such data processing needs as a critical opportunity to apply its Akida technology, which specializes in Event-Based Neuromorphic processing of ViT, CNN, TENN, and RNN. This paper will explore the need for efficient edge AI for time-series data and the startup BrainChip's ability to perform well at low power in this emerging market.

Sequential Data Analysis

Sequential analysis refers to the process of analyzing and extracting insights from data that is collected and organized in chronological order. This data type typically involves measurements or observations taken at regular intervals over time. Time-series analysis techniques aim to understand data patterns, trends, and dependencies and make predictions or forecasts based on historical patterns.

Here are a few use cases of sequential data analysis:

1. Financial Analysis: Time-series analysis is extensively used in finance to study stock market trends, analyze economic indicators, and forecast future market conditions. It helps model asset prices, trends, risk assessment, and portfolio optimization.
2. Demand Forecasting: Sequential analysis is crucial in demand forecasting for retail, supply chain management, and manufacturing industries. By analyzing

historical sales or demand data, businesses can predict future demand patterns and optimize their production, inventory, and supply chain accordingly.

3. Predictive Maintenance: Predictive maintenance can monitor equipment and machinery in real-time. Analyzing sensor data and historical patterns can help detect anomalies and predict potential failures, enabling proactive maintenance and minimizing downtime.
4. Energy Consumption Analysis: Utilities and energy companies use time-series analysis to analyze energy consumption patterns, identify peak demand periods, and optimize energy generation and distribution. It aids in load forecasting, energy pricing, and demand-side management.
5. IoT Sensor Data Analysis: With the proliferation of Internet of Things (IoT) devices, time-series analysis is widely used to analyze sensor data. It helps monitor and control smart homes, smart cities, environmental monitoring, and industrial processes.

The market size for applying AI in time-series data analysis is continually growing as organizations recognize the value of extracting insights and making accurate predictions from temporal data. While specific market size figures for this realm are not readily available, the broader AI market, including applications in time-series analysis, is expected to grow substantially. According to a report by Grand View Research, the global AI market size was valued at USD 62.35 billion in 2020 and is projected to expand at a compound annual growth rate (CAGR) of 40.2% from 2021 to 2028. This growth encompasses various AI applications, including time-series analysis, across multiple industries.

Adding the wrinkle of Time into Neural Networks

Traditional Convolutional Neural Networks (CNNs) have been around for 30+ years and combine multiple hidden layers trained in a supervised manner. These are sequential in nature and hence referred to as feed-forward neural networks. Bi-directional networks, also called Recurrent Neural Networks (RNNs), invented at the turn of the century, added capability for more complex learning such as language modeling. But for applications to time series, a machine learning engineer needed a combination of CNNs and a temporal network for spatial-temporal analysis. While academia developed networks that did temporal convolution, none have been power efficient or easy to train to make it to the far Edge.

TENNs: Temporal Event-based Neural Networks

Enter Temporal Event Based Neural Nets (TENNs). BrainChip, the first company to commercialize neuromorphic or event-based processing IP, has extended this to efficiently combine spatial and temporal convolutions to process sequential data in an innovative approach. A significant benefit is that it overcomes the training complexity and trains just like the simpler CNNs, but with the added benefit of reducing models' size without losing accuracy. All of this leads to improved performance and greater efficiency in the execution of complex models, which is imperative for Edge AI devices.

The advantages of using TENNs (Temporal Event-based Neural Networks) for analysis include:

- They learn to represent the temporal structure of the data, which can be essential for tasks such as forecasting and anomaly detection.
- They can make predictions for future time steps.
- They can be trained on large datasets of time series data.

Overall, TENNs are a powerful tool for processing time series data. They can learn to represent the temporal structure of the data and make predictions for future time steps.

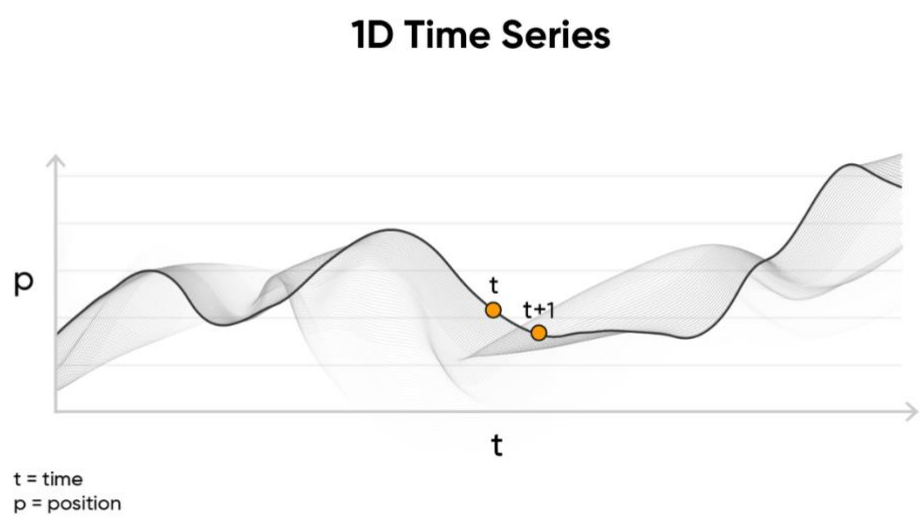

Figure 1: Traditional time series analysis applies to all kinds of signals.

Beyond the traditional Edge time series applications like predictive maintenance and anomaly detection, Brainchip claims their TENNs architecture can minimize or eliminate the need for DSP filtering of raw audio signals and vital signs in healthcare monitoring, which if it delivers on the promise, could mean very compact solutions for wearables, hearables or even implantable devices that sip energy in microjoules for each inference. A radical improvement for preventative healthcare.

But TENNs go further and start treating streaming inputs like video like a time series of frames, performing a 3D convolution comprising of a temporal convolution on the time axis and a spatial convolution on the XY axis. The secret Brain chip claims is the efficient way they achieve this convolution, enabling more advanced, higher resolution video object detection in tens of milliwatts.

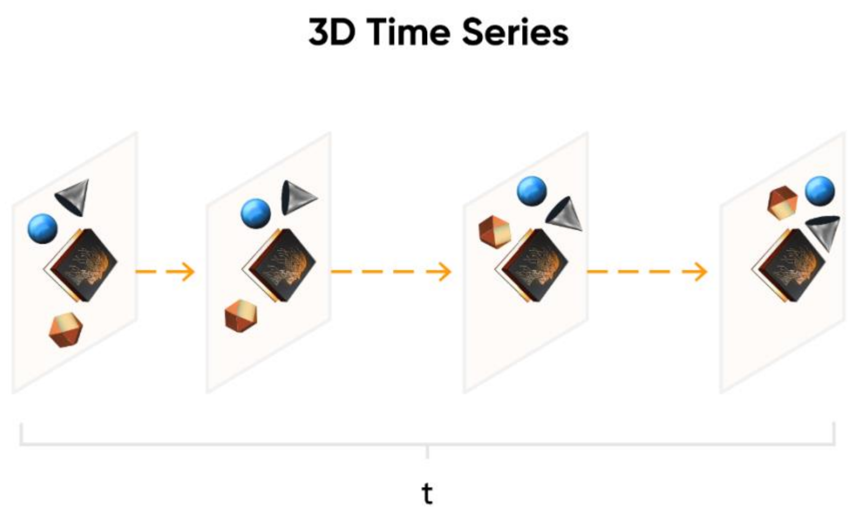

Figure 2: Treating streaming inputs as a 3D time series of frames with a highly efficient 3D convolution.

TENNs can be configured to operate in a buffered temporal convolution or recurrent mode. This flexibility allows developers to adapt the network to the specific requirements of their applications. Furthermore, TENNs benefit from efficient training on parallel hardware, potentially in the cloud with GPUs and TPUs like convolutional networks, while maintaining the compactness of RNNs for efficient inference at the Edge, whittling down the exponentially growing cost of training, which is a consistent concern.

The Akida 2nd generation Processor and TENNs

The BrainChip Akida processor is inspired by the energy-efficient way of the human brain's functionality. Akida, unlike traditional neuromorphic approaches, which are analog, is a fully digital portable processor IP that can perform tasks such as image classification, semantic segmentation, and odor recognition, including time-series analysis. It supports most current neural network architectures in addition to TENNs.

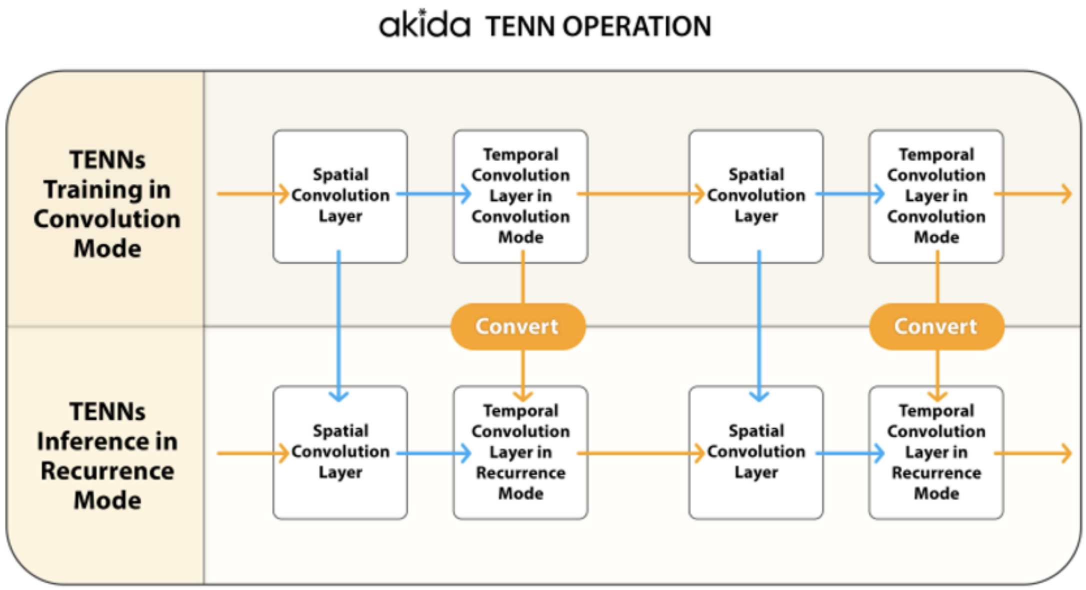

Figure 3: The TENN operation may combine transform from convolutions in training to recurrence operation for inference.

The Akida processor uses unique, highly parallel event-based neural processing cores. It has innovatively merged neuromorphic processing with native support for more traditional convolutional capabilities and functions, including hardware support for TENN networks. Its neuromorphic processing cores communicate using sparse, asynchronous events. This event-based processing method is well-suited for time-series data analysis because it efficiently manages high-speed, asynchronous, and continuous data streams.

BrainChip's Akida brings this innovative ability to look at vision, video, and other three-dimensional data as time series. An object visible through multiple two-dimensional frames, computed with the time element as the third dimension, makes video object detection much more interesting. Akida's support for efficient spatial-temporal convolutions makes this use case significantly faster with lower energy consumption.

The foundations of the Akida Processor.

The processor features a high-speed, low-power digital design optimized for edge computing applications, enabling real-time processing and low-latency analysis. Akida can process structured and unstructured data and learn and recognize patterns from streaming data, which is necessary for time-series data analysis.

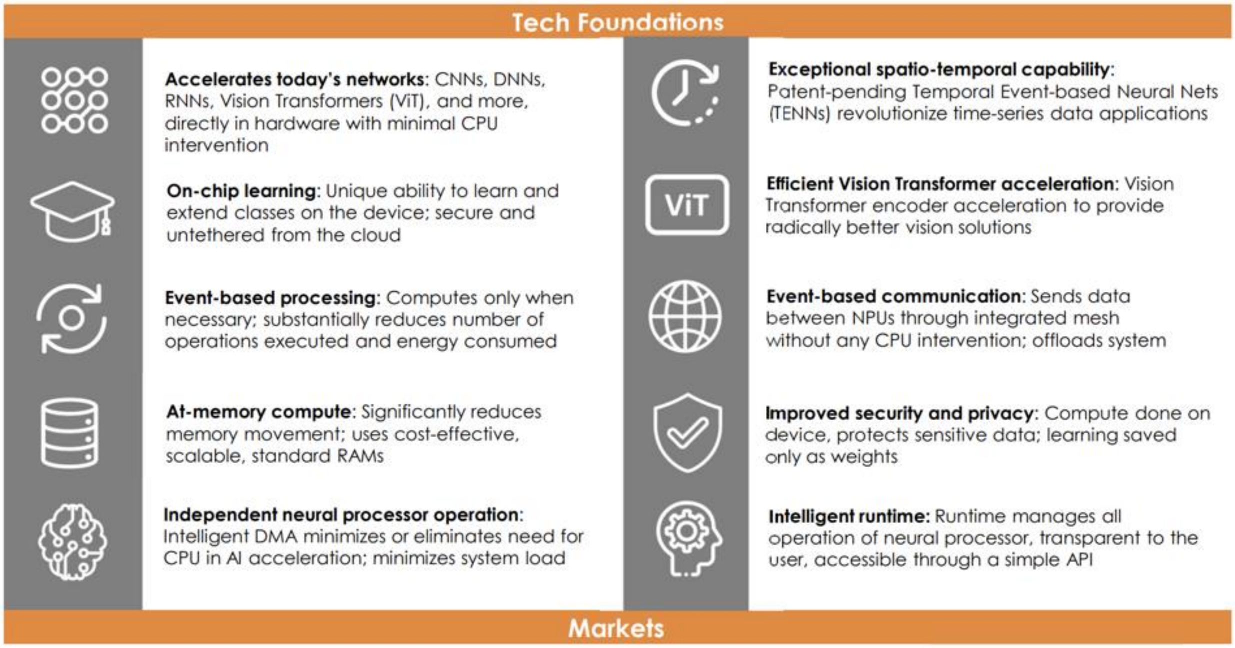

Figure 4: The foundations of the Akida architecture.

Akida provides a traditional CNN accelerator and adds TENN and Vision Transformer logic for a more comprehensive solution to sequential processing. The Akida processor is particularly effective for real-time data classification, anomaly detection, and predictive analytics. Consequently, the second-generation Akida processor, with TENNs support, is designed to provide efficient and accelerated hardware solutions for not just traditional one-dimensional (1D) time-series analysis tasks and various types of signals but takes the time-series paradigm to multi-dimensional applications like video object detection and vision using the underlying event-based processing paradigms

Conclusions

TENNs have demonstrated state-of-the-art performance across various domains of sequential data, as highlighted in BrainChip's recent white paper, "Temporal Event-based Neural Networks: A New Approach to Temporal Processing." Notable achievements include Raw Audio Speech Classification on the 10-Class Speech Classification SC10 dataset, Vital Signs Prediction on the BIDMC dataset, 2D Object Detection on the KITTI Vision Benchmark Suite (frame-based camera video), and 2D Object Detection on the Event-Based Prophesee 1 Megapixel Automotive Detection

Dataset. TENNs offer superior performance with a fraction of the computational requirements and significantly fewer parameters than other network architectures. This efficiency makes them an elegant solution for highly accurate models that support video and time series data at the Edge. It looks extremely attractive on raw signal data, which it can consume directly without needing DSP/filtering, making exceptionally compact audio management applications, including denoising. The MetaTF tools that plug into existing frameworks like TensorFlow and formats like ONNX simplify model evaluation, development, and optimization. (MetaTF is a free download from the BrainChip website at https://brainchip.com/developer/.)

The second-generation Akida processor IP is available now from BrainChip for inclusion in any SoC and comes complete with a software stack tuned for this unique architecture. We encourage companies to investigate this technology, especially those implementing time series or sequential data applications. Given that GenAI and LLMs generally involve sequence prediction, and advances made for pre-trained language models for event-based architectures with SpikeGPT, the compactness and capabilities of BrainChip's TENNs and the availability of Vision Transformer in second-generation Akida could facilitate more GenAI capabilities at the Edge.

IMPORTANT INFORMATION ABOUT THIS PAPER

Contributors and Publisher

Karl Freund, Founder and Principal Analyst, Cambrian-AI Research LLC

Inquiries

Contact us if you want 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 cited in context, displaying author's name, author's title, and "Cambrian-AI Research". Non-press and non-analysts must receive prior written permission by Cambrian-AI Research for any citations.

Licensing

Disclosures

This document was developed with BrainChip Inc. funding and support. Although the document may utilize publicly available material from various vendors, 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 opinions expressed herein are subject to change without notice.