Smart Inference Devices

The Wave of Perceptive Electronics Powered by Machine Learning

Sponsored by Perceive

Simon Solotko
Senior Media/XR Analyst, TIRIAS Research
simon@tiriasresearch.com

March 2020

Contents

Summary

The future of consumer device innovation lies in creating a new internet of smarter devices. This future will be powered by more perceptive sensors capable of local machine learning inference. Utilizing these sensors and multiple inference networks concurrently will drive advancement in virtually all aspects of smart device functionality and user experience. Privacy can be improved by running inference locally, with only the most deliberate transmission of user data and sensor feeds to the cloud. Perceive Ergo is a new inference processor designed for small devices delivering a 10X improvement in performance per watt over today's world class inference processors, with the potential to bring high accuracy, data center class inference to small, low power, high volume consumer devices.

Introduction – Machine Learning is Locked in the Cloud

The intersection of machine learning and smart devices promises to unlock a new wave of innovation in consumer electronics. Yet there is a very large gap between the processing requirements of the best machine learning networks and the performance of low power processors. The solution today is to utilize devices to gather and broadcast sensor data to the cloud, where high-power, dedicated machine learning processors can run inference and upon completion return outcomes back over the Internet to the user device. This approach provides machine learning capability but comes with significant disadvantages. Devices must spend their power budget on persistent network connections. The latency of cloud computing limits the utility of inference to the device and can break the user experience. The requirement to send raw data makes devices hard to secure and creates privacy concerns. Together, these limit the practical utility of machine learning for smart devices.

The first generation of low power machine learning processors lack the overall capacity and compute horsepower to handle all but the most basic networks with applications focused on speech command recognition and feature detection for camera control and custom filters. Machine learning core logic is largely gated by a slowing Moore's law - without a dramatic improvement in performance machine learning will have to remain in the cloud. Only a breakthrough in compute architecture will create devices capable of high performance, high accuracy local inference.

Perceive has demonstrated a new inference processor designed for small, low power devices. Unlike all previous inference processors, Perceive Ergo brings a level of performance to small devices previously possible only in powerful cloud-based inference processors. Ergo delivers over 4 GPU-equivalent floating-point TOPS peak performance at less than a tenth of a watt peak power. Efficiency is an unprecedented 55 TOPS/watt, a full order of magnitude better than any existing inference processor. It is designed to run multiple, large neural networks in excess of 100 million weights and network size exceeding 400MB with some variance depending on the implemented inference networks. Today, Perceive Ergo runs speech, facial detection, and object detection inference all concurrently using state of the art machine learning implementations.

Inference previously only possible in the data center can now be introduced to low power, high volume devices enabling inference driven device designs.

Tirias Research accepted the opportunity to write an independent white paper on the technology and applications of smart inference devices at the network edge sponsored by Perceive. This paper is intended to look to a future – now significantly closer – where even complex inference networks can be run in virtually any consumer device. The new wave of "smart inference devices" will provide high performance machine learning locally, keeping sensitive user data off the network and the cloud. They will be autonomous, utilizing machine learning to improve both low level device functionality and user experiences. They will be driven by rapid innovation, harnessing simultaneous progress in machine learning, processor design, and device design to create the breakthrough user experiences of the future.

Inference Breakthroughs

The advanced challenges recently addressed by machine learning show great promise for emerging smart inference devices. The vast number of teams contributing to machine learning research and learning/inference code bases has placed machine learning itself on an exponential learning curve. While most advancements are ultimately relevant to devices through cloud connectivity or local processing, those applicable to focused tasks with smaller datasets are the best candidate for purely local processing. Recent innovations with direct applicability include:

Enabling Devices to Observe Gestures and Manipulate Real World Objects

Advancements have been made in two related areas – tracking human and robotic hands, and teaching robots to perform tactile manipulations. In 2019, finger dexterity was demonstrated by OpenAI in a robot trained in 3D simulation that transfers its knowledge to reality, adapting to real-world physics. This created the flexibility to conduct complex tasks – like solving Rubik's cube in the real world – without real world training. Further, visual sensor hand tracking was introduced on the Oculus Quest in 2019 utilizing four cameras simultaneously required for 6DoF head tracking. The solution provides skeletal tracking plus several gestures utilizing 500mw on a Snapdragon 835 with a 3MB neural network. Previous non-ML solutions utilized depth cameras and dedicated logic at a total power of more than 15W with significantly lower accuracy

Enabling Devices to Weigh Future Consequences of Present Actions

Reinforcement learning was modified in Google's DeepMind to understand the long-term consequences of decisions in games. Temporal Value Transport was employed to return lessons from the future to inform the present, incorporating the probability of future benefits into present actions. The ability to incorporate future consequences into current decisions points broadly to improved real world decision making for machine learning systems. Devices capable of more complex decision making will vastly improve the number possible of tasks and complexity of applications.

Smart Inference Devices

Employing inference to orchestrate device functions is a powerful paradigm for future device design. Inference driven design can create a new class of user experiences while improving low level device functionality. Devices powered by multi-network inference will be equipped for complex interpretations of user and environmental input without engaging the cloud. The use of multiple concurrent networks will enable devices to achieve an advanced understanding and response to user and environmental input. The impacted logic of these devices will include:

Smart Inference Device Activation

Machine learning is well suited to interpreting voice, gestures, and visual input to activate devices, conserving battery power and reducing cloud data transfer.

Measured Cloud Utilization & Data Privacy

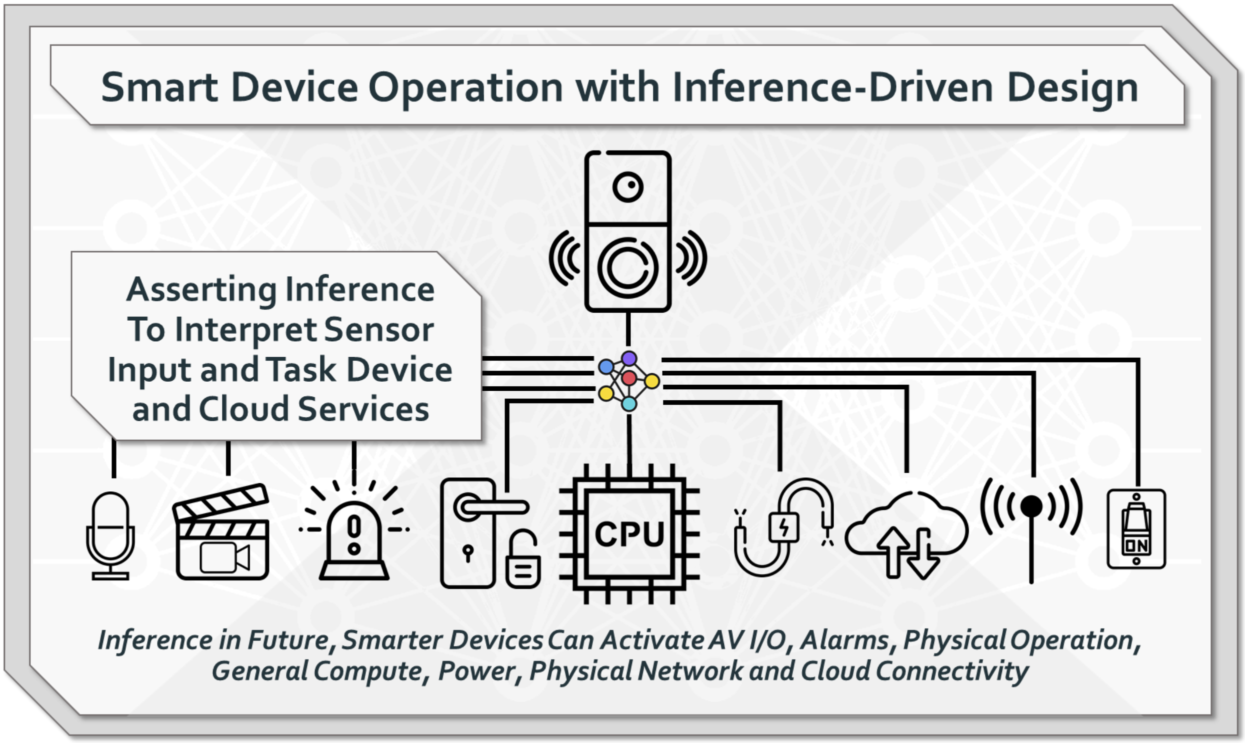
Inference can be employed by devices for tasking of the cloud including further inference or non-inference cloud computing. Devices able to make inferences locally do not need to provide raw data to the cloud. Voice, imagery, and sensor data, and the resulting inferences never need to be online. Devices can arbitrate which data is sent to the cloud, and the logic can be designed to enhance privacy and data security rather than to openly expose it due to a requirement to process it in the cloud.

Inference Driven User Experience

Inference has proven unmatched in providing device functions ranging from device dexterity, realistic voice interpretation and reproduction, navigation, vision, detection, identification and so on. Inference can be employed directly to create new-user level capabilities and experiences simply unavailable on devices incapable of inference. Processors capable of utilizing more than one neural network concurrently have the potential to utilize visual, audio, and innovative sensor input simultaneously to make complex decisions at the device and user level.

Smart Inference Device Activation

Activating sensors or connecting to the cloud incur computation, power consumption, and cost. If a device can rapidly infer when a sensor input requires subsequent action, processors can be intelligently activated, wireless connections established, and the cloud can be purposefully tasked. In this way low power inference can save overall device power by keeping high frequency tasks like interesting motion detection, voice activation, and user intent local to the device.


Smart Device Operation with Inference-Driven Design
Asserting Inference To Interpret Sensor Input and Task Device and Cloud Services
CPU
ON
Inference in Future, Smarter Devices Can Activate AV I/O, Alarms, Physical Operation, General Compute, Power, Physical Network and Cloud Connectivity

The importance of smart activation should not be underestimated – devices can move into acceptable power ranges and cloud services can avoid double digit false positives in many applications.

- Smarter activation words that only prompt device and cloud activity when tasking phrases and intonation are present
- Smart motion sensing that divines the intent of a motion to avoid downstream processing in a broad range of camera-based applications
- Smart sensor processing offloading general purpose processors and implementing ML to intelligently activate devices

Measured Cloud Utilization & Data Privacy

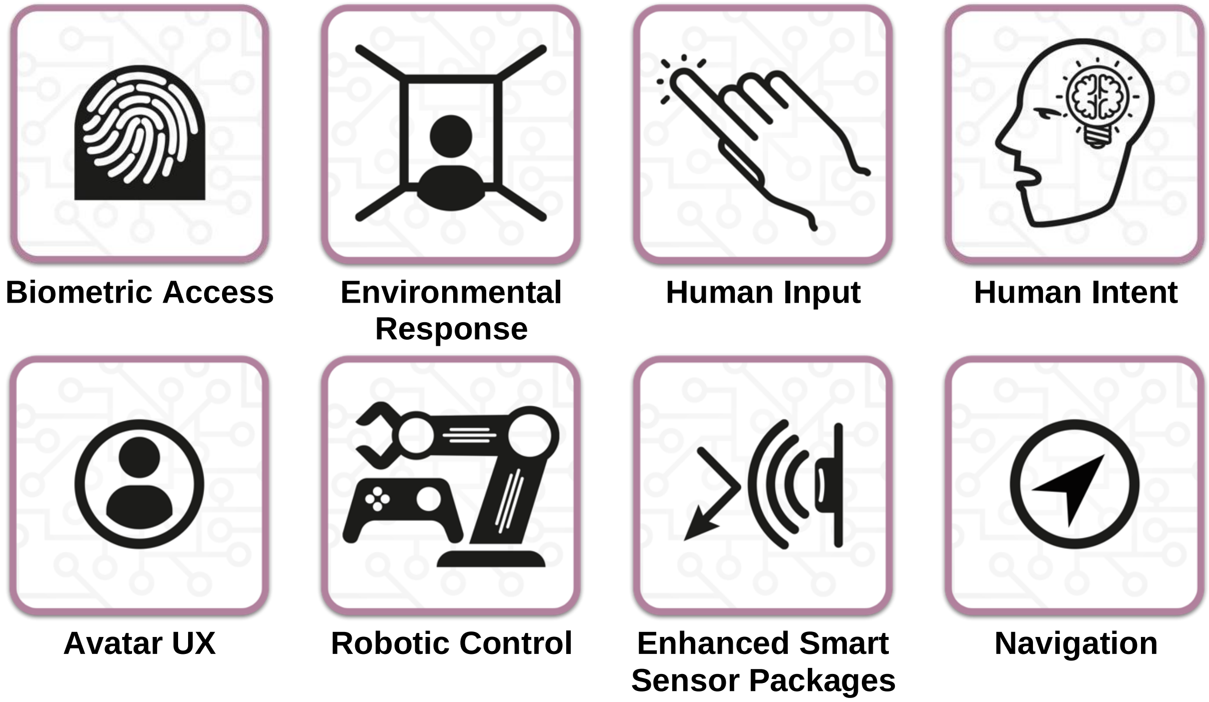
Mobile and social applications create compute and privacy challenges at unprecedented scale. Applications being designed for the cloud can migrate inference to the device, decreasing latency and saving cloud resources. Furthermore, the device – such as a wearable or smart speaker - can arbitrate local, multi-factor queues to task different, even multi-vendor cloud services. While the cloud may have computational horsepower and vast data, local devices that do not rely on network connections are free from network latency and unreliability. Devices able to accurately run inference locally can send fewer feeds to the cloud by analyzing audio, video, spatial, and other sensor data locally vastly improving opportunities to protect user privacy.

- Low latency user response achieved by shifting ML from the cloud to the local device can vastly improve user experience for voice, camera, and gesture inputs

- Local video and sound analysis can be analyzed locally, sending that video or audio upstream only when cloud-based analysis is required thus protecting sensitive user raw data streams
- Smart app interfaces can allow apps from multiple vendors to be tasked and invoke cloud interactions from a single smart device

Inference Driven User Experience

Inference as a core capability unlocks sophisticated and new user level capabilities in devices. The combination of machine learning inference for audio and visual processing allows devices to discern complex commands and contexts, and then produce sophisticated outcomes. Capabilities that can span emerging smart inference devices include user responsive capabilities like biometric access, environmental response, intentional input and derived human intent. Capabilities include avatar emulation, robotic control, smart sensor control, and navigation.



Desirable Attributes of a Smart Inference Device Machine Processor

A processor capable of orchestrating smart device function would utilize machine learning while containing critical interfaces to device sensors, core processing, and power control. It would become the mind that activated device function and arbitrated critical processes. Ideally, it would deliver functionality useful to making devices better products by reducing power consumption, intelligently activating device functions when required, and ensuring they were invoked according to an orchestrated need. At the same time, it would be capable of accurately processing complex networks with high efficiency and a small form factor – ready for insertion into affordable devices with desirable form factors.

A critical consideration in future smart devices is the relation between sensors and power consumption. Many devices require battery operation for extended periods. Today, complex visual sensor data analysis can consume all the compute resources at maximum TDP. Smart sensors will activate processing, wireless networks, cloud interaction, and physical function only when required. Smart sensors and device power reduction are required to accelerate smart home adoption. Intelligently activating network operation and cloud interactions can reduce device power consumption, extend operating times, and increase power budgets available for sophisticated functionality. Home wiring is a longstanding limitation, requiring many devices to be battery operated to smooth consumer adoption. Today's battery powered smart home devices often use up to 4AAs or rechargeable 3500mAh batteries and target operation times are 3+ months. Utilizing a smart sensor to activate devices only when required, and processing inputs locally without wireless network connectivity, can significantly increase capability while diminishing required power.

The intersection of optimized performance, architecture, software, power, and development platform are all required for inference to become a driving factor in the design of future smart devices. An optimal inference processor will excel simultaneously across these attributes.

Performance

- High ML Network Performance/Watt: Able to run high accuracy, contemporary video/image/voice processing network types
- Consistent Acceleration: Relatively uniform acceleration even with network code changes and spanning multiple network types

Architecture

- Integrated Design: Low footprint and high integration-simplifying design
- Standard Bus: Easy integration into devices using standard interfaces and bus design
- Multi-Sensor I/O: Support for video, audio, common & emerging sensors
- Scalability: Architecture able to scale to multiple ASICs or larger, higher performance ASICs with the same code
- Low Latency: Integrated memory architecture and cache for rapid processing of ingested data including audio and high resolution/multi camera video

Software

- Fast Load: Quickly and dynamically load and execute multiple ML networks
- ML Network Agnostic: Execute any neural network style or layer type
- Run & Correlate Multiple ML Processes: Run multiple networks and perform analysis spanning multiple inferences

- General Purpose Operation: Execute code to produce fully formed outcomes relevant to the device operation

Power

- Low TDP: Support battery power for extended use of wearables and disconnected operation
- Low Thermal Envelope: Low thermals ideally supporting wearables & fully passive operation
- Fast Power Up: Fast, low latency accelerator startup from user or sensor cues, ideally sub-frame and below user perception levels
- Selective Power Up: Command and control of power up of sensors and supplemental processing to minimize device average power consumption

Platform

- Strong SDK & Documentation: Software toolset makes it easy to deploy to the target accelerator ideally with 3rd party tool support
- Optimized Prefabs: Drop-in support for major application building blocks like detection, classification, denoise, etc.
- Deployment Toolset: Strong support and tool-based porting from major ML platforms e.g. MXNet, PyTorch, Caffe & TensorFlow

Perceive Ergo

Perceive has introduced Ergo, a fully integrated inference processor designed to offload all inference processing in low power applications and small footprint devices. Ergo has the ability to run inference with an equivalent performance per watt of over 55 TOPS/W and 4 TOPS at full power without sacrificing accuracy or limiting the kinds of networks that can be supported. The Ergo ASIC is packaged in a 7x7mm FBGA and processing many networks in ~20mW, with a maximum power of ~120mW, and fully passive cooling. In live demonstrations Ergo runs cool to the touch under full load.

Ergo is designed to run networks traditionally only possible on datacenter class inference processors. Today Ergo runs full YOLOv3 with 64M parameters at 246fps with a batch size of 1. Ergo can execute networks traditionally requiring in excess of 400MB of storage and over 100M parameters.

Compared to prior inference processors targeting low power applications, Ergo targets and achieves 20X to 100X the power efficiency making all prior processors and dedicated accelerators appear to have more or less equivalent performance per watt. Inference processors today are generally below 5 TOPS/W where Ergo stands out at 55+ TOPS/W.

To achieve this performance, Perceive has developed a novel compute architecture that retains high accuracy but vastly reduces memory and power requirements. Ergo's novel network representation circumvents the need for an array of MACs for inference and, further, is compact enough to run even large networks entirely within on-chip memory. The Ergo chip also employs aggressive power and clock gating for increased power efficiency. Thus, Ergo can provide extremely high accuracy within a 7mmx7mm package. This combination of a mathematically principled approach to ML, an architecture not based on MACs, no external memory, and traditional power-saving techniques is what gives Ergo its high accuracy, performance, and efficiency on data center-class networks.

The broad range of ported networks and the consistency of performance uplift indicate the company has been successful in creating an architecture capable of delivering a significant improvement in performance relevant to today's inference workloads. In addition, the company has demonstrated many multi-network implementations consistent with their network capacity and performance claims.

Perceive Ergo is designed to directly interface to high resolution, high frame rate video sensors with the opportunity for multi-sensor and real time metadata as additional inputs to inference processing. This creates the opportunity for advanced problem solving and multi-network inference, which can be employed for core device control and advanced end-user features. With high performance and network capacity, truly novel capabilities should be possible. With this opportunity comes the challenge of software design and training, presenting new technical challenges for device makers. Perceive has sought to make this development easier with a toolkit that includes ready-to-deploy networks for common machine learning applications.

Perceive Ergo Machine Learning Network Examples

Perceive Ergo can run multiple networks concurrently allowing smart devices to adopt inference driven design. It has been tested with contemporary multi-layer networks including CNNs (including residual edges), LSTMs, and RNNs and others. Demonstrated networks include:

Multi-Object Detection with M2Det

M2Det (Multi-Level Multi-Scale Detector) is a recent network (January 2019) for object detection and localization designed to detect objects of widely differing scales. M2Det is an end-to-end, single-shot object detector which is useful in real world applications where objects can be of radically different size and proportion in an evolving scene.

Multi-Object Detection with YOLOv3

YOLO is a CNN created by Joseph Redmon and Ali Farhadi that identifies and locates up to 80 object types in images and videos. Today YOLOv3 is one of the most popular multi-object detectors in datacenters.

Audio Event Detection with Proprietary Network

Optimized network able to discern multiple audio event classes with small network size making it ideal to employ in combination with larger visual processing networks.

Face Recognition with ResNet

Deep residual learning is noted for easier training and excellent accuracy introduced by Microsoft Research in 2015. ResNet in multiple layer configurations has been employed for large sample local facial and image recognition.

Perceive Ergo Concurrent Inference Network Examples

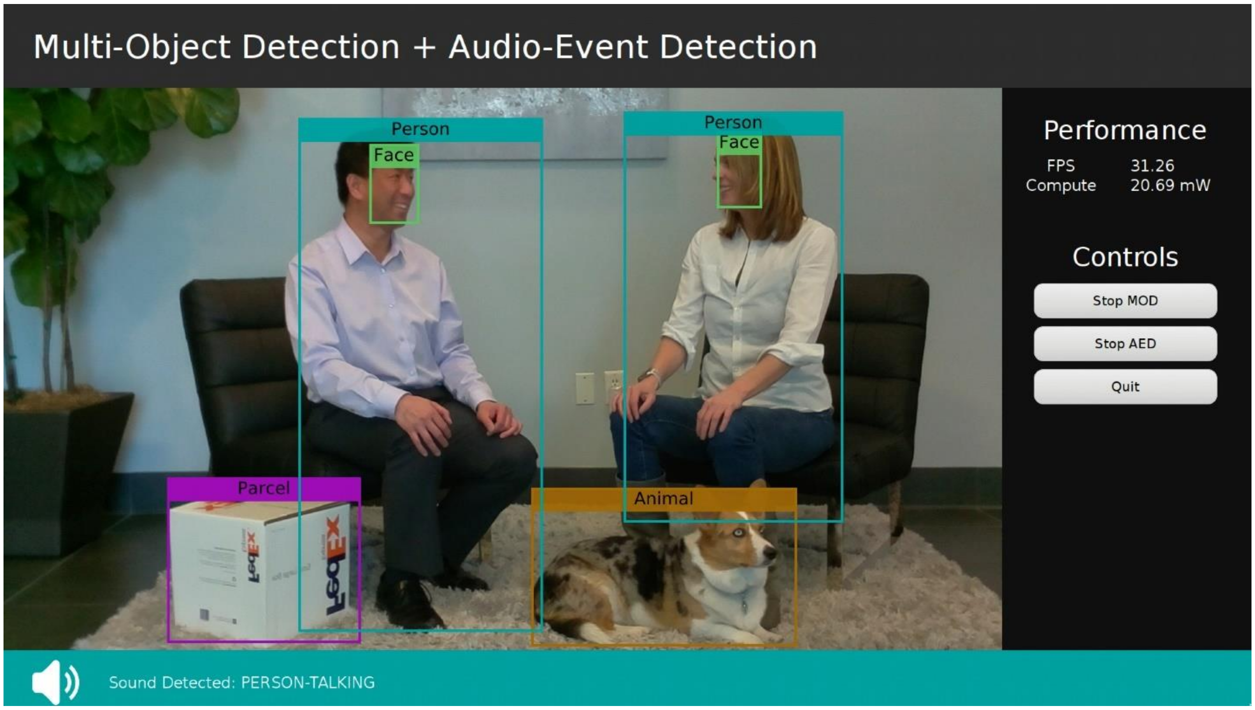
Perceive Ergo has been demonstrated on combinations of these networks and is technically capable of running multiple networks concurrently within its memory / network weight capacity. The processor is able to run novel combinations of networks with data from multiple sensors utilizing the onboard I/Os.

Perceive Ergo Complex Multi-Object Type Visual Detection and Recognition

Perceive Ergo was demonstrated concurrently running M2Det, Proprietary Face Feature Detection, and Resnet28 Face Recognition to simultaneously detect objects and identify persons from a high definition video source. The demonstration runs M2Det (73M weights), a proprietary network for face feature detection (0.5M weights), and Resnet28 face recognition (11M weights). Combinations of visual inference can comprehend and drive interactions and complex decision trees.

Perceive Ergo Concurrent Audio & Video Inference Demonstration

In this demonstration two styles of networks are run concurrently. Multi-object video detection using M2Det with 73M weights detecting 5 classes: person, face, animal, package, and vehicle (not pictured). Audio event detection using a proprietary network with 0.7M weights detecting 3 classes: person talking, smoke alarm (not pictured), CO alarm (not pictured). Combinations of both visual and audio inference can be employed to provide user interfaces and passive alerts concurrently with visual environmental context.


Multi-Object Detection + Audio-Event Detection

Performance

FPS 31.26
Compute 20.69 mW

Controls

Sound Detected: PERSON-TALKING

A live demonstration where Perceive Ergo performs concurrent audio & visual inference with M2Det detecting 5 classes from video input and a proprietary network detecting 3 classes for audio where inference consumes ~20mW. Courtesy Perceive.

The First Wave of Smart Inference Consumer Devices

The potential for smarter devices to assist in everyday life will be unlocked by advancements in training and inference. Imagining the breadth of everyday application categories and specific devices is powerful in designing forward looking product roadmaps.

Smart Locks & Entry Cameras: Security and entry control provided by smart locks, doorbells, gates, and security cameras. Designs must accommodate battery operation and extended operating time. Utilization of ML for video analysis of identity and intent, voice activated control, selective device power network and cloud access.

Smart Appliances: The rise of appliance makers with deep roots in electronics has driven innovation and competition. Smart sensors looking inside and outside of appliances creates opportunities for traditional smart home as well as improved overall functionality and automation.

Consumer Mobile Cameras: Visual processing as machine learning promises to provide next generation cameras features including low light photography, night vision, advanced visual processing, real time object identification/tagging/tracking, and automated turn on/recording in the presence of objects of interest.

Consumer Wearables & Portable Electronics: Devices with advanced user interfaces or visual processing, and emerging smartphone class wearables will benefit from previously unavailable levels of real time processing and cloud independent functionality. Capabilities include smart analysis of sensor data for health, home, and hobbies.

Consumer Smartphones: Smartphones have become de-facto compute devices but today rely on the cloud for higher order machine learning. Today's basic machine learning applications – biometrics and imaging – will be replaced with more intelligent agents which utilize the cloud selectively lowering latency and decreasing service costs.

Consumer Robotics: Home robots are mobile devices which might support larger batteries but have mechanical and compute functions competing for power. Low power inference can improve physical control and local autonomy without sending sensitive visual data to the cloud.

Consumer/Light Commercial Drones: Increases in resolution, and a desire for automated feature detection and navigation, are driving requirements for high performance, low power machine learning. Logic built around feature detection can guide subsequent actions, alerts, and notifications while driving complex mapping, surveillance, agriculture, and preservation applications.

AR Head Mounted Displays: AR is challenged by the requirement of low power and low heat while performing the complex visual processing critical to augmented reality. Challenging AR machine learning use cases require a significant investment in an integrated software and hardware platform. However, in time, AR/VR are seen to be a major driver of consumer machine learning.

The New Internet of Smarter Things

Making a really useful and indispensable device, it turns out, is extremely hard to do. Today's Internet of Things depends on devices utilizing persistent connections to tap inference services in the cloud. This complexity makes the deployment of inference an expensive, high power affair. Connected devices have suffered from challenges including accuracy, power, complexity, cost, processing performance, network connectivity requirements, user interface design and so on. It may be for these reasons that the Internet of Things has seen a rise and fall in expectations - consumers still live in largely unconnected homes, largely ignore wearable tech, and fear their personal privacy is lost to the Internet.

Local inference may be able to improve device function on enough fronts to propel the Internet of Things back to relevance. Smart devices that can accurately infer locally have the promise to dramatically improve fundamental device operation while expanding capabilities and increasing user privacy. Inference led design may simplify device logic and, by keeping the cloud out of the loop for basic operations, could dramatically improve the reliability of critical functions like user interface and much improved baseline/disconnected capabilities. There have been references to the “AI of Things” however we believe we – the purveyors of high tech - should not get ahead of ourselves. We are nowhere near AI and the Internet of Things has already suffered enough hype. Creating better devices requires pragmatism and would benefit from less over selling and more over delivering. Simply smarter devices that employ advanced inference capabilities may bring enough reliable capability in smaller, lower power form factors to win over users and reassert the role of the consumer electronics device.

Perception in Smart Devices

Moving inference out of the cloud and into everyday devices creates exciting opportunities for innovation in consumer electronics. Local inference can orchestrate device activation, cloud utilization, and multiple simultaneous neural networks to deliver new, valuable user experiences.

Perceive Ergo is the first processor to place the emerging best-in-class machine learning research within the reach of smart devices. By interfacing directly to sensors, high fidelity input can be interpreted, without arbitration by the CPU, at extremely low power. In this way, inference processing can be employed to initiate complex tasking of other compute, electrical and mechanical functions of a smart device. This new solution architecture – inference driven design - is a challenge to device makers, who now have a surprising amount of machine learning horsepower available to drive features that were probably several years out on the roadmap. While challenging, the opportunity to create the first wave of smart inference devices is the most disruptive and exciting opportunity available in consumer electronics today.

The information contained in this report was believed to be reliable when written, but is not guaranteed as to its accuracy or completeness.

The contents of this report represent the interpretation and analysis of statistics and information that is either generally available to the public or released by responsible agencies or individuals.

This report shall be treated at all times as a confidential and proprietary document for internal use only of TIRIAS Research clients who are the original subscriber to this report. TIRIAS Research reserves the right to cancel your subscription or contract in full if its information is copied or distributed to other divisions of the subscribing company without the prior written approval of TIRIAS Research.