State of the Field:

Artificial Intelligence in Embedded Systems

Pranav Chainani

Table of Contents

Brains Behind the Machine

Personality is to the human body as artificial intelligence is to an embedded system. Artificial intelligence (AI), the buzz phrase that has been growing in familiarity as a supposed solution to creating more efficient way of living. Artificial Intelligence refers to the idea of machines mimicking humans' ability to make data driven decisions. Companies are currently improving the user experience of their products through the integration of AI to help with personalized recommendations, user profiling, and customer service.

Recent advancements in increasing computational power, data availability, and machine learning algorithms have allowed for tasks that typically require human intelligence to be performed faster, and more efficiently at a much larger scale as it analyzes vast amounts of data in real time. The rapidly growing trust in AI's decision-making ability has been an encouraging sign for technological innovation. AI hopefuls believe that its scope of responsibility will soon include early disease detection, energy optimization, and personalized education just to name a few.

To make these dreams a reality, innovators work towards a common goal: **maximize task efficiency with minimal space, power, and resources**. Humans will constantly keep pushing the limits by making technology operate with the same efficiency in a smaller space. That is where integrating artificial intelligence into embedded systems comes into fruition. An embedded system is a specialized tiny computer built into a device to perform specific tasks.

However, the race to achieve maximal task completion within small pieces of hardware has several obstacles. High computation requires high energy usage, embedded systems need extended battery life to run AI algorithms. The need to compute data in real-time is challenging for small devices as they tend to have memory constraints for their small microcontrollers and microprocessors. This has spurred companies to create their own specialized chips.

This paper will provide an overview of how artificial intelligence is currently being integrated into embedded systems. It will cover important phrases and concepts that help the understanding of how it is made possible, why it is necessary, and key considerations to keep in mind as innovation rapidly progresses.

Background

What are Embedded Systems?

To further understand what enables artificial intelligence software to operate on small devices, it is imperative to study the common make up of an embedded system [1].

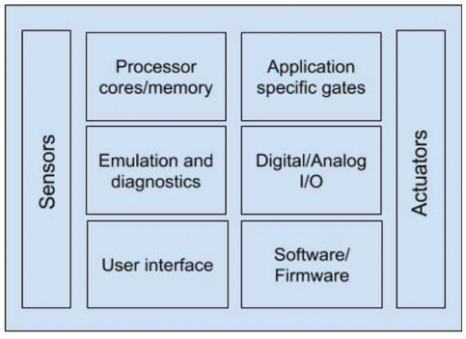

Figure 1: Embedded System Architecture [1]

Processor (CPU): Carries out instructions and calculations to make the system function.
Random Access Memory (RAM): Temporary memory storage.
Sensors: Extract information from the environment.
Actuators: Interact with the environment.
Digital Analog I/O: Receives input and produces output.
Software/Firmware: Controls the behavior of the device to perform specific tasks.

These embedded systems are found integrated in items to increase efficiency and reduce the need for constant human maintenance. Here are some examples of how everyday items utilize embedded systems to make life easier:

- Digital Camera: Uses sensors to capture light and turn it to digital signals.
- Washing Machine: Uses sensors to detect water levels, load managements.
- Thermostat: Processor analyzes sensor data to adjust heating/cooling levels.

How do we make them smarter?

Embedded systems help make objects "smart." AI makes them even smarter. Having machines perform tasks that typically need human decision making, learning, and reasoning requires increased computational power. To understand this increase we must explore what makes artificial intelligence possible.

AI has a history that dates to the mid 20th century where mathematicians such as Alan Turing, and John McCarthy studied concepts such as symbolic reasoning and artificial neural networks [9].

Here is a brief history of how AI has evolved [10]:

- Turing's 1950 paper, "Computing Machinery and Intelligence," proposed the Turing test as a measure of machine intelligence.
- John McCarthy coined the term "Artificial Intelligence," in 1956 during the Dartmouth conference.
- Initial hype led to philosophical debates such as the "Chinese Room" argument that questioned whether machines could think or only simulate understanding.
- A series of chess programs were written in the 1970s to simulate ideal play to win tournaments.
- IBM's Deep Blue program defeated chess champion, Gary Kasparov, in 1997.
- In the 21st century, applications evolved to healthcare, finance, and autonomous systems as companies realized the broad applications of data driven decision-making.

Transformers – An AI Breakthrough

In 2017, researchers at Google created a paper title "Attention is All You Need." This presented the architecture of a transformer which revolutionized **natural language processing** [15]. They are a type of AI model that can process large amounts of text through focusing on the relationships between words in a sentence. The paper highlights a technique called **self-attention** to decipher which words are the most important for understanding context. This laid the groundwork for large language models such as ChatGPT.

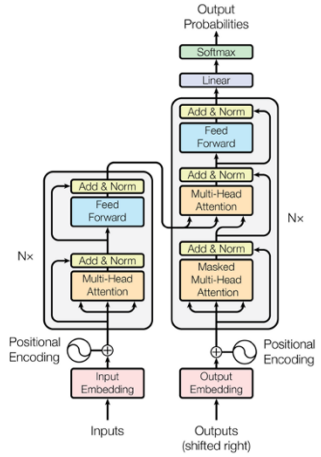

Figure 2: The Transformer Model [17]

Key terms:

- **Machine Learning** - Machines interpreting data to improve its performance.

- **Internet of Things (IoT)** - A network of devices that communicate and share information.
- **Neural Network** – A computational model inspired by the human brain structure.
- **Natural Language Processing (NLP):** Enables machines to understand and interpret human language.

However, improvements to software are only as good as the hardware that can do justice to it by not hindering the expected performance. That is why the current wave of innovation is focused on making embedded systems compatible with complex software.

Current Innovations

Hardware that incorporates artificial intelligence will have a high impact on whatever industry it is applied in as system efficiency can always be improved. Here are a few current innovations and applied fields.

Figure 3: Autonomous Robotic Surgery Prototype [18]

Healthcare

Transformative potential has been demonstrated with AI in embedded systems used for safety-critical applications like Autonomous robotic surgery is a large hurdle for mankind due to its many safety considerations. AI powered systems such as the **Smart Tissue Autonomous Robot (STAR)** use **deep learning (DL)** and **Convolutional Neural Networks (CNNs)** to perform complex surgical tasks with increased accuracy [4]. This involves tasks like suturing and vascular access with precision.

Patient safety considerations are the top priority for such research. These systems incorporate **Safety Integrity Levels (SIL)** and explainable AI (XAI) frameworks to provide greater system transparency [4]. Techniques such as transfer learning are utilized to help adapt pre-trained models for unseen surgical scenarios. However, there are safe mechanisms and robust anomaly detection algorithms to prevent performance from degrading.

The distant goal is full autonomy. Limitations for achieving this is due to the difficulty of finding safe ways to test such technology without putting humans' lives at risk. However, the advancements described depict the potential for achieving a high-level of autonomy that would be overseen by humans. This would be where robots independently execute surgical plans with minimal intervention ensuring reliability and efficiency during operations.

Edge AI & Tiny ML

Edge AI refers to the "edge" of the network such as IoT devices, sensors, and servers where data is retrieved and analyzed. Incorporating Edge AI into embedded systems would allow for devices to run AI algorithms locally without having to rely on the cloud for computing. This enables real-time calculations, reduced latency, and data privacy. Deploying AI at the edge requires techniques like **model quantization** [13]. This is where neural network weights are reduced from 32-bit integers to 8-bit integers to decrease model size and increase inference speed.

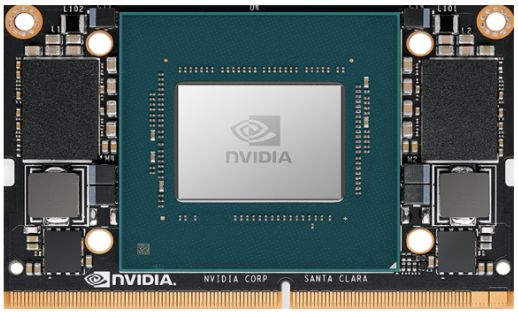

Figure 4: Nvidia's Jetson Xavier AI Supercomputer for Embedded Systems [8]

Edge AI utilizes hardware advancements such as modern energy-efficient processors like Nvidia's Jetson Graphics Processing Unit (GPU) and Google's Tensor Processing Units (TPU). Leveraging the parallelism of GPUs and TPUs is important to achieving high level performance. For example, Google's TPU accelerates task inferences while Nvidia's Jetson can provide scalable computing power that is correlated to the increased complexity of tasks [13].

Tiny Machine Learning revolves around using machine learning on tiny and low-power devices such as microcontrollers and IoT sensors [14]. The integration of machine learning algorithms on these devices allows for decisions to be made locally. While TinyML and Edge AI share many of the same underlying technologies, TinyML focuses more on ultra-low-power devices with resource constraints while Edge AI focuses on more capable devices.

More aggressive quantization and customized neural networks are needed to operate on small devices. A large application of TinyML is personalized healthcare. This is due to the resource constraints and need for on-device computation of wearable technology. Additionally, the sensitivity of health data is protected by TinyML's avoidance of sending raw data to external servers.

Cameras

An example of how the integration of AI has improved the embedded system of the camera is the *Intelligent Camera* patent [11]. It represents a significant advancement in imaging technology as it combines AI, multi-sensor inputs, and specialized hardware to deliver real-time video and image advancements. The utilization of neural networks and depth sensors allows for it to perform tasks with precision. These tasks include object segmentation, background blurring, and aesthetic optimization.

The camera has been able to run complex software through the integration of advanced hardware like TPUs and GPUs. This enables the system to achieve processing times of less than a millisecond for dynamic background replacement, lighting adjustments, and real-time framing.

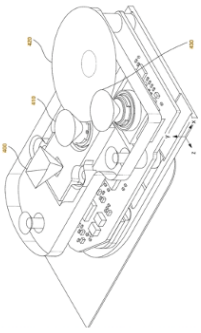

Figure 5: Intelligent Camera Design [11]

Applications for such a camera include enhancing cinematic videography, security surveillance, augmented reality, and smartphone photography.

Energy Efficiency

Energy usage optimization calls for real-time sensing and adjustment. The *AI Energy Usage Sensor* [12] is a system that businesses can leverage to combat this issue. It uses deep neural networks and advanced machine learning algorithms to interpret the data that is collected from sensors within a building's power systems. Through breaking down single energy signals, it identifies individual appliances and their power consumption levels.

The unique aspect of this system lies in its ability to perform **real-time demand forecasting** as well as **automated demand management**. The algorithm can predict peak energy usage through understanding historical energy patterns with varying weather and geographical data. After outlining these forecasts, it proactively adjusts power consumption and schedules device operation to mitigate the peak levels.

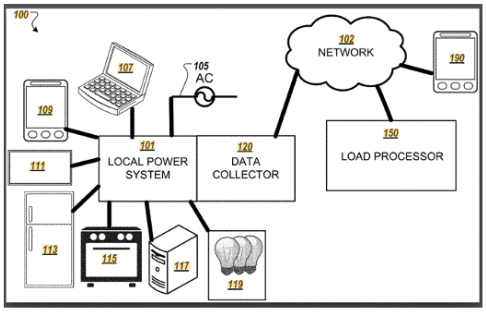

Figure 6: AI Energy Usage Sensor Process [12]

This demonstrates how AI within embedded systems is optimizing solutions for energy efficiency and reducing operational costs for building managers.

Issues with Current Technology

As people and companies demand more complex AI models and systems, challenges emerge that tend to negatively impact their development and adoption. These issues involve technical and ethical domains that require attention to ensure a responsible deployment of AI in embedded systems.

Power Consumption

The demand for real-time, high-performance AI calculations strains the limited power and space of embedded systems [3]. Techniques such as **Dynamic Voltage Scaling (DVS)** and **Frequency Scaling (DFS)** has been utilized to reduce energy usage. DVS adjusts the processor's supply voltage to meet performance requirements for a specific workload. DFS adjusts the processor's clock frequency during periods of inactivity or low processing demand which ensures that energy consumption aligns with the needs.

However, there are critical drawbacks with these methods. A significant issue is latency [7]. Switching voltage and frequency in real-time introduces delays that disrupt the prompt execution of tasks. This is important when considering systems where timeliness is key as seen in transportation and medical applications.

Additionally, constant changes in voltage and frequency pose the issue of necessary thermal management. These changes can lead to uneven heat distribution which degrades hardware reliability, particularly in nanoscale CMOS (Complementary Metal-Oxide Semiconductor) structures that are very common in modern electronics [3].

Gathering Data

To achieve the level of computation to make noticeable differences in society, lots of data is necessary to train AI models to reach high levels of accuracy. Attempting to gather data such as consumer behavior and medical images is costly and can be viewed as an invasion of privacy. However, the static nature of AI knowledge systems does not allow for the systems to learn from experience unless programmed to do so [10].

Ethics and Safety

AI's decision-making capabilities are a product of the humans that create it. They require humans to embed values to prevent harm and bias. However, this leads to the AI being influenced with whatever bias the creator possesses. Issues such as biased algorithms, ethical dilemmas in autonomous vehicles, and opinion-based responses pose as safety threats to the user.

Key ethical challenges in AI for embedded systems include [5]:

- Bias
- Discrimination
- Transparency
- Privacy Concerns
- Autonomy vs Control

With idealistic visions of AI being able to replace a portion of human decision-making, a series of safety considerations persist:

- **Data Security:** Collecting vast amounts of sensitive data are imperative to improving the system's performance. Therefore, concerns have been raised pertaining to data breaches and legislation that acts as a safeguard.
- **Reliability:** Relying on AI systems for actionable outcomes could potentially cause physical and mental harm on humans. The chance of lacking necessary reasoning that a proactive human possesses could be dangerous making it difficult to test AI solutions for healthcare applications [4].
- **Biased/Unbiased Decision Making:** Biases are necessary at times where human discretion is used. An objective AI system may make decisions that go against societal morals based on the community is applied in.

Next Steps

Notable Efforts to advance the innovation and address the issues of artificial intelligence in embedded systems can be depicted through the following:

Neuromorphic Computing

Through offering brain-inspired hardware solutions to mimic neural architecture, neuromorphic computing is a groundbreaking advancement that could solve issues related to power consumption and complex computation. Traditional AI architectures like CPUs and GPUs consume a great amount of power due to the separation of computation and memory, requiring frequent transfers of data. Neuromorphic systems combine computation and memory to reduce latency and energy consumption. An example of this is Intel's Loihi chip with 130,000 neurons and 130 million synapses that allows for real-time learning and **parallel computation** [16].

These systems run only when an event occurs compared to traditional architecture that runs on a clock frequency, saving energy [16].

The Loihi chip can execute some workloads 10,000 times more efficiently than conventional processors as well as execute demanding workloads up to 1000 times quicker. Since the chip is still in the research and development phase, it lacks compatibility to be adopted to consumer use cases.

Human Cyber Physical Intelligence (HCPI)

A transformative approach to the design of an embedded system that prioritizes a symbiotic relationship between humans and AI systems. This addresses the issue of ethics and safety as it promotes a more dynamic relationship between human and machinery.

These systems are improving their data gathering issues as they can be used to mirror human learning processes through using **Case Based Reasoning (CBR)** where humans reference solutions to past problems to make the systems adaptive [6]. Embedded systems would be utilized to monitor human's physical, psychological, and thinking activities through various sensors and then can make educated decisions to drive efficiency [2].

The system approach involves three layers:

- *Situation Awareness Layer*: Integrates data from social and physical spaces.
- *Cognitive Processing Layer*: Utilizes data to understand and optimize situations.
- *Target Decision-Making Layer*: Makes decisions to drive the best operation of the system.

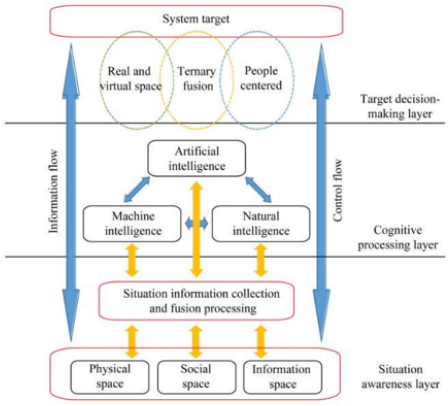

Figure 7: Human Cyber Physical Intelligence Hierarchy [6]

Conclusion

Artificial intelligence needs a compatible vessel that will allow it to thrive. Humans have been constantly exploring to maintain and improve their own vessel so their objectives can be fulfilled. We are now in a state of self-reflection. We attempt to incite the optimal ways in which we learn, think, and make decisions into an inanimate object. While justification for how the brain operates remains largely unsolved, we are constantly attempting to find ways to automate tasks we find tedious or make up for areas we lack precision in. Artificial intelligence within embedded systems is an exciting field that is enabling humans to be creative through automating responsibilities and presenting data efficiently. Its ability to be trained on different types of data allows for systems to cater specifically to the pain points of humans. Cameras taking optimal photos, surgeries performed flawlessly, and environmental disasters accurately forecasted are all ideas that now seem obtainable. Improvements in software such as Edge AI and Tiny ML are enabling quick computation times while helping the issue of data privacy. Improvements in hardware such as Nvidia and Google's AI chips are what bring the artificial intelligence models to life as power consumption must be managed.

However, it is easy to get caught up in the marvels of this burning flame. The repercussions of unsafe AI can be immensely harmful if not deployed with proper safeguards. People yearn for answers and if the credibility of AI is perceived to be superior it can lead to

uninformed decision-making and physical harm. As this becomes a more comfortable and knowledgeable field, more discussions around ethics, legislation, and safeguards will allow AI to be safely deployed into embedded systems that will make our lives easier.

Bibliography

[1] D. Chikurtev, S. Bogdanov, N. Spasova, and V. Ivaniv, "Prerequisites for a self-sustaining embedded system with Artificial Intelligence," 2020 XXIX International Scientific Conference Electronics (ET), pp. 1–4, Sep. 2020. Doi:10.1109/et50336.2020.9238328

[2] L. Cao and R. Chen, "Enhancing tennis online teaching through Big Data and artificial intelligence in Embedded Systems," Computer-Aided Design and Applications, pp. 70–82, Sep. 2023. doi:10.14733/cadaps.2024.s8.70-82

[3] N. Ganesan and T. Muthumanickam, "Artificial Intelligence System based embedded real-time system power optimization and adaptability," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), pp. 1–6, Nov. 2022. doi:10.1109/icaiss55157.2022.10011023

[4] S. Sophia and K. Markus, "Artificial Intelligence in safety-relevant embedded systems - on autonomous robotic surgery," 2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 506–509, Jul. 2021. doi:10.1109/iiai-aai53430.2021.00089

[5] V. Vakkuri and P. Abrahamsson, "The key concepts of ethics of Artificial Intelligence," 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), pp. 1–6, Jun. 2018. doi:10.1109/ice.2018.8436265

[6] X. Li, "The new development direction of Artificial Intelligence—human cyber physical intelligence," 2021 10th International Conference on Educational and Information Technology (ICEIT), pp. 249–252, Jan. 2021. doi:10.1109/iceit51700.2021.9375623

[7] Y. Zhang, "Research on embedded speech recognition system based on Artificial Intelligence," 2021 IEEE International Conference on Industrial Application of Artificial Intelligence (IAAI), pp. 215–220, Dec. 2021. doi:10.1109/iaai54625.2021.9699877

[8] "The World's smallest AI supercomputer," NVIDIA, https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-nx/ (accessed Nov. 19, 2024).

[9] "Libguides: Artificial Intelligence (A.I.): History of artificial intelligence," History of Artificial Intelligence - Artificial Intelligence (A.I.) - LibGuides at Illinois Central College, https://library.icc.edu/c.php?g=1372140&p=10141462 (accessed Nov. 27, 2024).

[10] C. Smith, T. Huang, G. Yang, and B. McGuire, The history of Artificial Intelligence - University of Washington, https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf (accessed Nov. 28, 2024).

[11] X. Miao and A. E. Rubin, "Intelligent Camera," Jul. 2, 2019

[12] D. Serven and J. Kvam, "System and methods for power system forecasting using deep neural networks," Mar. 14, 2023

[13] S. S. Gill *et al.*, "Edge Ai: A taxonomy, systematic review and Future Directions," *Cluster Computing*, vol. 28, no. 1, Oct. 2024. doi:10.1007/s10586-024-04686-y

[14] J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, and S. Han, "Tiny Machine Learning: Progress and Futures, "*IEEE Circuits and Systems Magazine*, vol. 23, no. 3, pp. 8–34, 2023. doi:10.1109/mcas.2023.3302182

[15] Md. Al-Amin, M. S. Ali, A. Salam, A. Khan, and A. Ali, "History of generative Artificial Intelligence (AI) chatbots: past, present, and future development," *arXiv*, Feb. 2024. doi:10.48550

[16] Y. S. Yang and Y. Kim, "Recent trend of neuromorphic computing hardware: Intel's Neuromorphic System Perspective," *2020 International SoC Design Conference (ISOCC)*, pp. 218–219, Oct. 2020. doi:10.1109/isocc50952.2020.9332961

[17] S. Cristina, "The Transformer model," MachineLearningMastery.com, https://machinelearningmastery.com/the-transformer-model/ (accessed Dec. 4, 2024).

[18] O. Barnes, "Robots give surgeons a helping hand," Financial Times, https://www.ft.com/content/2c47aaba-29e3-4f6a-b1ef-6037fa68513d (accessed Dec. 4, 2024).