TinyML devices enabling physical GPT
About us

the AI hardware partner

Seeed Studio has been a leading Open Hardware company since 2008, empowering half a million direct users to create real-world digital solutions. Through relentless efforts and earned trust, our ever-growing product lines now form around emerging AI scenarios:

- Sensor networks to fetch extensive real-time data
- Edge computing to push intelligence to new frontiers

We provide industrial-ready modules and devices, and open up the capability of prototype, produce, and promote as Fusion service. Innovators from different vertical domains co-create with us to make their creations widely available for diversified markets.

By embracing open source, community building and integrated software suites like SenseCraft, we are proactively lowering the tech barriers and including users with diverse expertise for glocalized matters.
From Technology to Industry

Technologies
- Open Source Hardware
- Machine Learning
- Advanced Sensors
- Home Assistant
- Wireless, DePIN
- LLM

Emerging Technologies
- Open Source Hardware
- Collaborative Innovation

Applications
- Asset Tracking
- Smart Building
- Industrial space
- Smart City
- Smart Agriculture
- Smart Energy
- Tech for good
- Open Science

Digital Economy
- Traditional Industries
- Smart Environment
- Smart Agriculture
- Smart Energy
Adding AI to almost anything

TinyML

- Faster
  - new architecture
    - Cortex-M55
- Cheaper
  - <10 $<1 w
- multi-modal
  - vision
  - sound
  - speech
  - sensors
- low power
- Easier
  - few shot training
  - no-code web server
Faster and Lower power

**Refresh Rate**

- 2.57 FPS
- 5.55 FPS
- 30.30 FPS

44 x

**Energy Efficiency**

- 3.79 watts
- 0.4 watts
- 0.35 watts

- **Raspberry Pi 4B**
- **Grove Vision AI v1**
- **Grove Vision AI v2** (M55 + U55 AI addon)

- **Grove Vision AI v1**
- **Grove Vision AI v2** (M55 + U55 AI addon)

- **XIAO ESP32S3**
- **Grove Vision AI v2** (M55 + U55 AI addon)
Faster

2024 MCU AI Vision Boards: Performance Comparison

How I conduct the test

1. flash the same test model on board - Swift-YOLO Tiny 96x96
2. feed the camera with the same human face picture under the same condition
3. record their performance

Notes:

1. Raspberry Pi 4B is included for perspective on CPU vs. MCU performance, despite its difference from MCU boards.
2. Nicla Vision: Due to compatibility issues with the test model, it is tested with an alternative method; results are not directly comparable with other boards.
3. The inclusion of these 2 boards is intended to offer a broader perspective on processing capabilities across different hardware platforms. Their results are color-highlighted for clear distinction.

Scan the code to read the article

*The larger the color block area, the better the combination.*
Smaller and cheaper

XIAO - tinyML MCUs
Add AI to Almost Everything

The Seeed Studio XIAO Series is a collection of thumb-sized, powerful microcontroller units (MCUs) tailor-made for space-conscious projects requiring high performance and wireless connectivity. Embodying the essence of popular hardware platforms such as ESP32, RP2040, nRF52840, and SAMD21, the Arduino compatible XIAO series is the perfect toolset for you to embrace tiny machine learning (tinyML) on the Edge. Trusted by 500,000 developers globally!

- **Module and Development Board Hybrid**
  Enabling Rapidly Prototype While Easily Integrate, Significantly Streamline Product Development Process

- **Invest for Your Future**
  Unified Form Factor Enables You to Seamlessly Upgrade or Downgrade Your Product at the Lowest Cost

- **Single-Sided Surface Mount Design**
  Effortlessly Incorporate XIAO into Other Boards for Large-Scale Manufacturing

- **Strong Ecosystem Support**
  Extensive Software Compatibility, Abundant Community Resources, and Dedicated Technical Assistance
SenseCraft Model Assistant is an AI platform dedicated to simplifying the training, distribution, and deployment of AI models. With just a few clicks, you can easily deploy models and say goodbye to tedious configuration and coding. It supports users to upload and share self-trained models, build a shared model library, and promote collaboration and innovation among AI enthusiasts. Currently supports computer vision algorithms (such as target detection, image classification, image segmentation, and pose) and LLM, making it possible to realize high-speed and accurate inference on low-cost hardware, unlocking the powerful potential of AI in edge devices.
Easier - No code model MLops

https://github.com/Seeed-Studio/Seeed_Arduino_SSCMA/tree/main/examples/camera_web_server
Easier - chat with TinyML

Message from Vision AI Alert: Human presence detected. Check the area.
Message from Vision AI Alert: Human presence detected. Check the area.
Message from Vision AI Alert: Human presence detected. Check the area.
Message from Vision AI Alert: Human presence detected. Check the area.
Message from Vision AI Alert: Human presence detected. Check the area.
Message from Vision AI Alert: Human presence detected. Check the area.
Message from Vision AI Alert: Human presence detected. Check the area.
Atlas of Sensors
Nerve endings of the digital world
what's next: few shot training?

https://huggingface.co/spaces/stevengrove/YOLO-World
Generative AI expand to real world

- Stronger
- Cheaper
- Distributed
- Specialized

- multi-modal
  - imagie
  - Speech
  - Sensors

- Locally deployed

Generative AI

- Cheaper
  - Nvidia Jetson Orin Nano 8GB from $599

- Specialized
  - RAG Agent

- > 40T
Evolving edge AI capability

AI Performance (TOPS)

- Entry-Level
- Vision AI
- Generative AI

**reComputer J1010**
NVIDIA Jetson Nano 4GB

**reServer J2032**
NVIDIA Jetson Xavier
NX 16GB

**reComputer J3011**
NVIDIA Jetson Orin
Nano 8GB

**reComputer J4011**
NVIDIA Jetson Orin
NX 8GB

**reServer Industrial J4012**
NVIDIA Jetson Orin
NX 16GB

**reServer AGX
Orin 350 Series**

*From $599*
Generative AI at the Edge

https://www.jetson-ai-lab.com/
## Cost comparison of Open AI vs Local LLM

### #5 Years Usage of Solution

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Times/day</th>
<th>Times for 5 Years</th>
<th>Jetson Solution Cost</th>
<th>OpenAI Solution Cost/yr</th>
<th>Number of months to break-even</th>
</tr>
</thead>
<tbody>
<tr>
<td>(24h) 1 request per day</td>
<td>1</td>
<td>1825</td>
<td>899</td>
<td>1.825</td>
<td>5993.33</td>
</tr>
<tr>
<td>(24h) 1 request per hour</td>
<td>24</td>
<td>43800</td>
<td>899</td>
<td>43.8</td>
<td>249.72</td>
</tr>
<tr>
<td>(24h) 1 request per 30min</td>
<td>48</td>
<td>87600</td>
<td>899</td>
<td>87.6</td>
<td>124.86</td>
</tr>
<tr>
<td>(24h) 1 request per 15min</td>
<td>99</td>
<td>180675</td>
<td>899</td>
<td>903.4</td>
<td>60.83</td>
</tr>
<tr>
<td>(24h) 1 request per 60s</td>
<td>1440</td>
<td>2628000</td>
<td>899</td>
<td>2628</td>
<td>4.16</td>
</tr>
<tr>
<td>(24h) 1 request per 10s</td>
<td>14400</td>
<td>26280000</td>
<td>899</td>
<td>26280</td>
<td>0.42</td>
</tr>
</tbody>
</table>

### Large Language Models

- **Cloud Service**
  - OpenAI
  - SenseCraft AI

- **Local Service**
  - NVIDIA® Jetson Orin

### Wired vs Wireless

- **Wi-Fi**

### Event Alarm

### Cons:
- Privacy
- High demand, lower cost with local LLM
- Local LLM speed is network-independent

### Pros:
- Low demand, lower cost with OpenAI
- OpenAI offers high accuracy
Next generation human machine interface to complex system

RAG and Home assistant integration

https://github.com/Seeed-Projects/LocalJARVIS
Convergence of TinyML and Generative AI

TinyML
- Faster
  - new architecture (Cortex-M55)
- low power
  - <1 w
- Cheaper
  - <10 $

Generative AI
- stronger
  - > 40T
- Distributed
  - Locally deployed
- Cheaper
  - Nvidia Jetson Orin Nano 8GB from $599
- Specialized
  - RAG Agent

Embodied AI
- Multi-modal
  - vision
  - sound
  - speech
  - sensors
- Next gen HMI
- Autonomous Machines
- Droids
- Mixed Reality

Multi-modal
- imagine
- Speech
- Sensors

Easier
- few shot training
- no-code
- web server
Convergence of TinyML and Generative AI

rhapsody of TinyML and local LLM
Physical AI agent for smarter facility and machines
Physical AI agent for smarter facility and machines

SenseCraft
Ready to use models

Image
Voice
Sensor data

tinyML for local detection

Large Language Models
Cloud Service
- OpenAI
- SenseCraft AI

Local Service
- NVIDIA® Jetson Orin

Event Alarm

Wi-Fi
Embodied AI as a bridge between science and application

- Meaningful Data
- Mixed Reality
- Autonomus Machines
- Droids
- next gen HMI

TinyML
- multi-modal
  - vision
  - sound
  - speech
  - sensors
- Faster
  - new architecture (Cortex-M55)
- low power
  - <1 w
- Cheaper
  - <10 $

Generative AI
- Distributed
  - Locally deployed
- Cheaper
  - Nvidia Jetson Orin Nano 8GB from $599
- Specialized
  - RAG Agent

Pervasive Insights
- Easier
  - few shot training
  - no-code web server

- Stronger
  - > 40T
Embodied AI as a bridge between science and application

IoT2wild Contest

Website: https://www.hackster.io/contests/iotinthewild
Winner announced at Hackster Impact Summit on October 11, 2022
Based on various digital transformation scenarios, we continue to develop smart devices that integrate the latest technologies, and work closely with developers and industry experts to provide software and hardware solutions for vertical industries at multiple levels.

Co-Invent Solutions

IoT Devices

seeed studio

Non-engineer Experts

Industrial Know-How

Software, Algorithm, Industrial Insights

Solution

Smart Agriculture

Smart Greenhouse

Energy Monitoring

Carbon Monitoring

Smart City

Smart Traffic
Let’s talk!
ep@seeed.cc