Unlocking the Power of Edge AI

White Paper

LatentAI
Adaptive AI for the Intelligent Edge

Latent AI, Inc.
info@latentai.com
Data is the new oil and we need data refineries along the edge continuum.
Unlocking the Power of Edge AI

by Sek Chai, CTO and Co-Founder, Latent AI

The scope of AI seems limitless. Business leaders and engineering teams are frequently considering compelling, intelligent applications powered by AI technology because AI is solving many business problems.

In this white paper, we are presenting key needs and solutions to enable the AI factory, which is essential to deploying the next generation of edge AI solutions.

The Case for the Edge Continuum

Data is the new oil, and we need data refineries along the edge continuum.

As companies integrate digital transformation around the world, they realize the value of AI technology. Raw data from devices, applications, users, and operators, can be harnessed to train AI models using a data-driven development process. The AI models are then used to automate, recommend, and perform other intelligent functions that make users and business operations more productive. AI-powered solutions such as intelligent transportation to drug discovery are here to stay.

The business imperative to extract value from the data is creating a bottleneck situation for the AI factory. The digital transformation has exacerbated the need for more inexpensive ways to manage and process the deluge of data. Large cloud-centric data lakes are no longer practical or scalable with the massive data sizes (terabytes per day) generated by sensors. Storing and processing data centrally in the cloud might not make economic or business sense for many industry-specific use cases.

LatentAI
The need for edge AI with processing near the source of the data is very clear. The shift for processing to the periphery helps avoid network overload, achieve user experience with timely responses, and allows businesses to scale up their AI technology rollout. There is no standard definition for edge AI, and the best solution can be a hybrid variant along the edge continuum. In our previous blogs, *AI Moves to the Edge* and *Exploring WebAssembly AI Services on Cloudflare Workers*, we highlight the need to consider the large spectrum of solutions from the sensor edge to the network edge.

**The AI Factory**

With new attention toward edge AI, organizations need to consider how to bring their AI models to run along the edge continuum in a systematic and automated manner.

All AI factories need a simplified AI development workflow that can be embedded into edge devices. AI model optimization is already challenging, and it is even more difficult for the edge because of constraints in memory, power, and processing capacity. Also, this workflow needs to be accomplished without legions of data scientists and developers working over many iterations of the same AI model before it can be deployed.

The AI factory workflow should address the following questions:

- Does it support different hardware targets?
- Can it handle state-of-art AI model architectures?
- Is it easy to use and automated?
The edge continuum is populated with heterogeneous devices from programmable (CPU, GPU) to hardware accelerators (FPGA, ASIC). New specialized processors are now entering the market that offers differentiated acceleration for edge inference. However, hardware vendor tools are precisely made for the vendor-specific silicon. These hardware-specific tool flows are often opaque to the type of AI transformations performed, and developers have no control over how their AI model accuracy is affected.

The ideal AI factory workflow should incorporate a homogenous development flow for all target hardware. It should be easy to use to aid the developer and help expedite the deployment of edge AI solutions. The workflow should be repeatable and robust to deliver compelling AI-driven solutions to the market.

What is MLOps?

MLOps is a shorthand for Machine Learning Operations, and it represents the set of best practices for the AI factory. It is a blend of AI development practices integrated into a traditional software development flow. MLOps includes engineering processes to efficiently and reliably deploy and maintain AI models into production MLOps focuses on improving the quality of the AI model through automation.

An AI factory with an integrated MLOps environment is the key to business success. MLOps enables the lifecycle of the engineering culture around AI system development and operations. System construction, such as testing, integration, deployment, infrastructure, and release management, are crucial elements of the
Edge AI poses additional challenges for MLOPs workflow. In addition to constraints in compute resources, MLOps workflow must address the asynchronous inference and the data sparsity in large-scale edge deployment. In layman's terms, edge AI inferences are not processed simultaneously or at the same physical location. Edge devices are not physically co-located, and the inferences may occur at different times. Consequently, proper orchestration is needed to derive deeper, more consistent analytics from edge AI deployments.

**The Latent AI Solution**

Latent AI is enabling the AI factory as businesses bring AI solutions to the edge continuum. With the LEIP SDK (Latent AI Efficient Inference Platform), AI practitioners have end-to-end developer tools that help extract the best performance of an AI model. LEIP is designed to take the hard work out of AI system development and deployment by making it easy for AI scientists, software developers, and IT/OT professionals to accelerate and solve real-world problems using edge AI.

The LEIP SDK is a fast, easy-to-use developer tool that builds upon existing AI/ML infrastructure and frameworks such as TensorFlow and Pytorch. The runtime engine generates an optimized binary that does not require additional libraries to run. The compiler can target various hardware (CPU, GPU, DSP) such as Intel x86, ARM processors, and Nvidia CUDA GPUs. Our quantization and compression algorithms can target an extensive range of target bit-precision, from floating-point to sub 8-bit integer precision. We have demonstrated state-of-the-art benchmark optimization results for [tinyML](https://tinyml.org) applications in keyword spotting (KWS), anomaly detection (AD), visual wake words (VWW), and tiny image classification (IC).
LEIP SDK is used in different phases of edge AI development:

- **Design & Explore** – In this first phase, developers use LEIP to estimate target model architecture and hardware. With support for different hardware, developers can estimate memory sizes and inference latency. LEIP helps reduce the overall solution cost with a more accurate bill of material (BOM).

- **Develop & Train** – In the second phase, developers use LEIP to train the edge AI model to reach the desired memory footprint and inference speed. LEIP excels in low bit-precision quantization below 8-bit integers. For example, this enables the reduction of a visual wakeup model to detect a person using only 81kB at 4-bit integer precision, with negligible accuracy loss.

- **Optimize & Deploy** – In the third phase, developers embed their models into target hardware to run at the fastest inference speeds. LEIP compiles and generates Latent Runtime Engines (LRE) for laptops (CPU/GPU) for testing and prototyping. When ready, the developer uses LEIP to generate run LRE for the target embedded processor.

LEIP provides vital benefits to the AI factory in many ways:

**Scaling**

LEIP is critical in scaling the AI factory’s number of AI models. By making development consistent across models and hardware targets, developers become more productive and can efficiently churn out AI solutions rapidly. Projects that take weeks are completed in just hours.
**Trust**

LEIP builds trust for managing machine learning across models and datasets by enabling a repeatable process through automation, testing, and validation. Developers can use the same test and validation software at all stages of development, enhancing the reliability and credibility of the AI model.

**Seamless Integration**

LEIP provides standardized and well-defined APIs to support integration into the CI/CD software development flow. The LEIP SDK provides clear handoff points between the data science and operations teams. This enhanced collaboration brings together teams with expertise in data to understand software deployment and maintenance. With LEIP, the AI factory becomes a highly adaptable production pipeline that can accommodate KPI-driven business goals.

**Better use of data**

LEIP enables the AI factory to capitalize on new data because, with shortened production cycles, the AI factory can improve the AI models with new iterations. This rapid development pace helps businesses drive reliable analytics and insights. LEIP helps reduce business risks with higher quality models using the latest data from real-world conditions.

**Model Lifecycle**

LEIP helps manage the AI model lifecycle, moving it from development to deployment. AI models have no business ROI until they are running in production. LEIP reduces business CAPEX with more flexible target hardware and reduces OPEX with faster development with less risk.
Conclusion

The codependency between AI and the edge is evident. AI needs the edge to get more responsive, and the edge needs AI to alleviate network bottlenecks. Businesses need AI factories that address the challenges of deployment on the edge continuum to enable edge AI. Latent AI's mission is to help these next-generation AI factories by providing developer tools that optimize neural network inferences for compute, energy, and memory while preserving algorithmic accuracy. The difference between success and failure is now based on how your AI factory scales to use their data. With Latent AI, you have the freedom to deploy your AI models wherever you need them to run.

About Latent AI

Latent AI takes the hard work out of bringing AI to edge at scale for any architecture, OS, framework or hardware target through our Adaptive AI technology. For more information about the availability of LEIP SDK, please feel free to contact us at info@latentai.com or check out our website.