3rd World Conference on
Engineering, Technology and Applied Science
November 18, 2024 | Bangkok, Thailand
https://doi.org/10.62422/978-81-974314-7-0-003

Unleashing Small Language Models (SLMs) Everywhere: For offline/on-device access - Edge and Web Deployment with Wasm & WebGPU and Mobile device deployment with Mediapipe

Nirav Kumar
Head of AI and Engineering, Navatech Group, Bangalore, Karnataka, India

In the rapidly evolving landscape of machine learning (ML) deployment, the demand for efficient access to models without extensive cloud infrastructure or constant internet connectivity is paramount. WebAssembly (Wasm) alongwith Mediapipe, are the technologies spearheading a new era of ML deployment. This talk delves into the power of leveraging Wasm and WebGPU along with projects like wasi-nn to deploy Small Language Models (SLMs) directly within web browsers and edge devices, reshaping the possibilities of on-device AI. We explore practical examples showcasing the fusion of Wasm's cross-platform execution capabilities and WebGPU's prowess in parallel computation, enabling developers to deploy SLMs seamlessly across diverse environments. These tools empower developers to harness the full potential of SLMs on the edge and web, providing them with the necessary infrastructure to deploy, optimize, and execute ML models efficiently in browser and edge environments.

Also deploying and accessing machine learning (ML) models (SLM's) over mobile devices using flutter efficiently poses significant challenges. Traditional methods rely on Native platforms and constant internet connectivity. This talk explores an approach for deploying Small Language Models (SLMs) directly within app thereby reducing reliance on constant internet access. Utilizing MediaPipe on flutter opens-up the opportunity to run SLM's models locally on device.

Biography:

Nirav Kumar: Leading Innovator in AI, Web, and App Development With a decade of experience in data science and machine learning and 15 years in web and app development, Nirav Kumar is a prominent leader in the tech world. As the Head of AI and Engineering at Navatech Group, he leads groundbreaking research and development projects aimed at advancing AI technology. Nirav has made notable contributions to the field of Applied AI, particularly in the realm of Conversational AI, making it accessible on web and mobile platforms.