lesson 1: Introduction to AI at the Edge

What is AI at the Edge?

The edge means local OR near local processing, as opposed to just anywhere in the cloud. This can be an actual local device like a smart watch with ECG capability and arythmia detection in built. The edge provide us :

- Low latency.
- real-time decision-making capabilities.

The edge AI algorithms can be trained on the cloud but they will not send any data to cloud when running, everything will be done locally

Application	Cloud vs Edge
Voice Assistant	Cloud
Self-Driving Cars	Edge
Insights from Millions of Sales Transactions	Cloud
Remote Nature Camera	Edge

Why is AI at the Edge Important?

Network impacts

Network communication can be expensive (bandwidth, power consumption, etc.) and sometimes impossible (think remote locations or during natural disasters)

Latency considerations

Real-time processing is necessary for applications, like self-driving cars, that can't handle latency in making important decisions

Security concerns

Edge applications could be using personal data (like health data) that could be sensitive if sent to cloud

Optimization for local inference

Optimization software, especially made for specific hardware, can help achieve great efficiency with edge AI models

---

Application of AI at the Edge

Applications

There are nearly endless possibilities with the edge some of them are:

- Self Driving Cars.
- Arythmia detection.
- Surgical Robots.
- Tracking objects.
- EEG based peripheral device (Brain computer Interface).
- Intruder Detection.
- CO and other harmful gas detection.

Historical Context

From the first network ATMs in the 1970's, to the World Wide Web in the 90's, and on up to smart meters in early 2000's, we've come a long way. From the constant use devices like phones to smart speakers, smart refrigerators, locks, warehouse applications and more, the IoT pool keeps expanding. oT growth has gone from 2 billion devices in 2006 to a projected 200 billion by 2020. Cloud computing has gotten a lot of the news in recent years, but the edge is also growing in importance.

Which of these are reasons for development of the Edge?

- Proliferation of Devices.
- Need for low-latency compute.
- Need for disconnected devices.

---

some things about OpenVINO toolkit

Flow of data in OpenVINO

- Train a model (Tensorflow, Caffe, MxNet etc)
- Run Model Optimizer
- IR format .xml & .bin files (Intermediate Representation)
- Inference Engine
- Edge Application

Distinct Topics

- Pre-trained models can be used to explore your options without the need to train a model. This pre-trained model can then be used with the Inference Engine, as it will already be in IR format. This can be integrated into your app and deployed at the edge.
- If you created your own model, or are leveraging a model not already in IR format (TensorFlow, PyTorch, Caffe, MXNet, etc), use the Model Optimizer first. This will then feed to the Inference Engine, which can be integrated into your app and deployed at the edge.
- While you'll be able to perform some amazingly efficient inference after feeding into the Inference Engine, you'll still want to appropriately handle the output for the edge application, and that's what we'll hit in the final lesson.

Summary

- The basics of the edge
- The importance of the edge and its history.
- Edge applications