Introduction to ONNX

Presented by Bhasker Raju

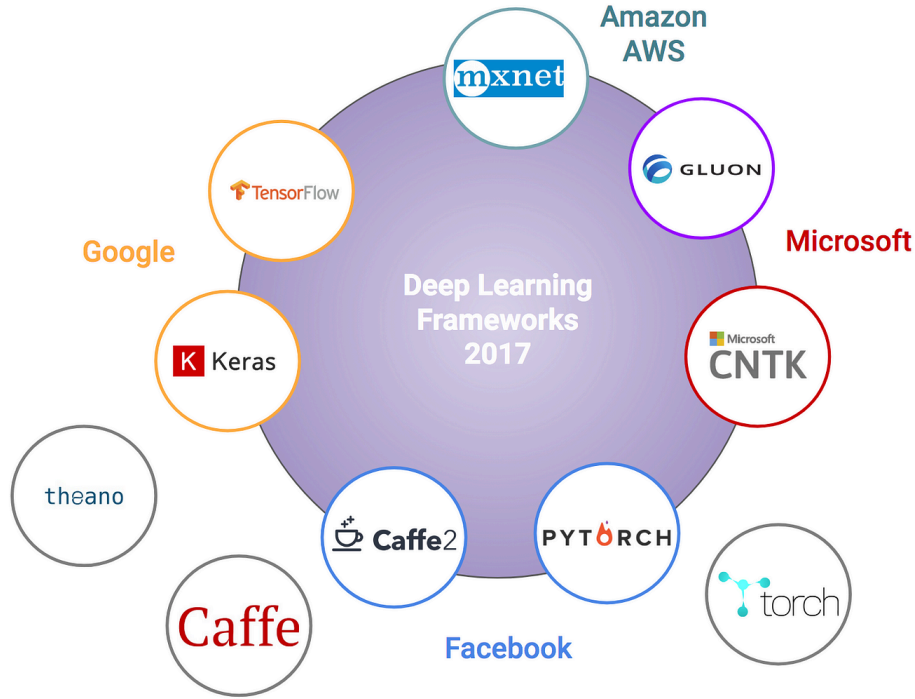
Agenda

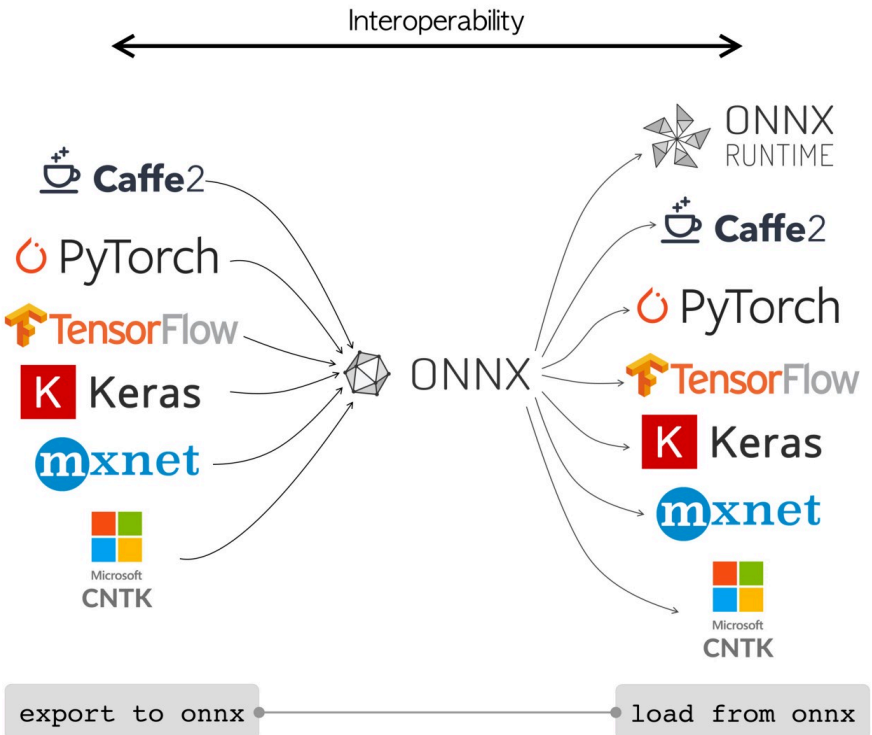
What you will be learning today

- 📝 **Concept** - What and why we need ONNX
- 👩‍💻 **ONNX Runtime** - Is ONNX framework agnostic ?
- 🤹 **Interaction** - where can we use ONNX?
- 🎥 **Demo** - Demo a python model in VS Code
- 🎥 **Demo** - Demo a GenAI model in browser
- 🛠️ **Architecture** - which technology powers ONNX

Pre-ONNX era

- Multiple frameworks from multiple vendors
- Diverse base Technology and Huge learning curve
- No strong developer community or support (except python)
- No cross compatibility between frameworks though written in same language.
- Difficult to collaborate multiple frameworks together. ex dotnet + python + cloud





The ONNX era

- Not a development framework but a universal ML format and has a runtime.
- Developed by `Microsoft` , `Facebook` and other developers.
- Doesnt matter in which framework or language a ML model is developed. You can convert to ONNX and use it.
- Runtime offers native execution speeds and interoperability.
- can run on any device and OS (linux, mac, windows, android, ios, raspberry, IoT etc)
- On-Device Training (advanced onnx feature)

Demo

ONNX runtime in VS Code

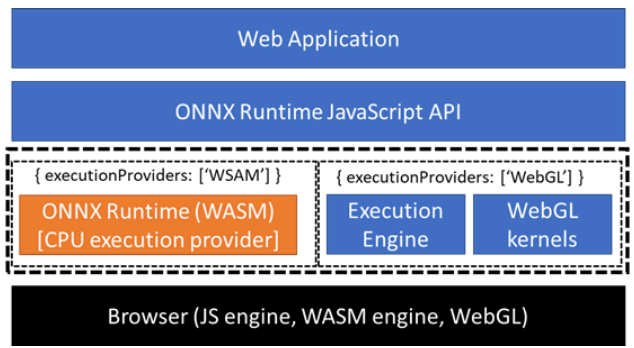
- Create a Linear regression (`y =mx+c`) in python using `pytorch` framework
- convert the model into `ONNX` format
- create a dotnet `console` app and import the `onnx` model into it.
- supply the input and get a prediction.
- import the same `onnx` model into `react` application to run in-browser.
- supply the value and get a prediction.

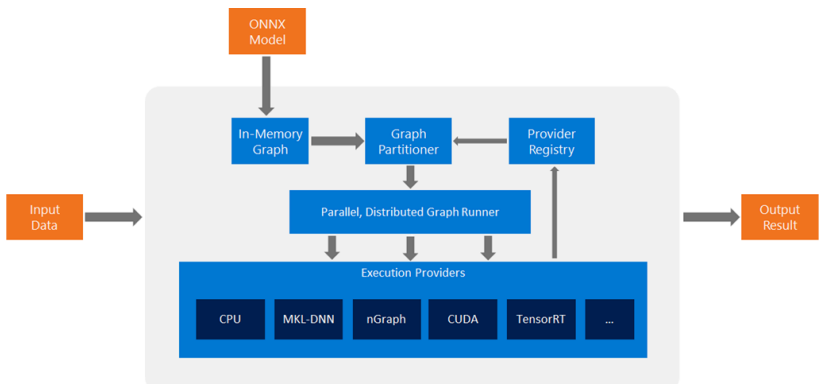
ONNX Architecture

Below is the architecture for browser and non-browsers

```
---
Browser
---
```

```
---
Non-Browser
---
```





Learn More

Documentation · GitHub . Phi 3.5 Vision .Phi3-Onnx .Onnx Zoo