Running an LLM on a Pi

Pete Warden, pete@usefulsensors.com

Why?

Large language models bring a lot of unique capabilities around interactions, question answering and knowledge retrieval.

Running them on the edge can help reduce costs, latency, and increase availability for locations with poor or no connectivity.

They’re also a lot of fun to play with!

What am I showing?

I'll demonstrate running a three-billion parameter LLM on a Raspberry Pi 5, in real time, with no additional hardware required.

This is part of the EE292D Edge ML class I teach at Stanford, and you can find full instructions at https://github.com/ee292d/labs/tree/main/lab1

Set Up

First, make sure you've gone through Lab 0 to set up your development environment.

Make sure your Pi can access all the memory it needs:

```
sudo su
echo "*               soft    memlock         unlimited" >> /etc/security/limits.conf
echo "*               hard    memlock         unlimited" >> /etc/security/limits.conf
reboot
```

Install GGML:

```
pip install --break-system-packages llama-cpp-python==0.1.77
```

Download the model:

```
./download_model.sh
```

Running the model

```
./run_model.py
```

```
Enter your question below:

> What is the capital of France?

 The capital of France is Paris.

> Are you sure?

 Yes, I am sure.

> Really? I thought France was a made up country!

 No, France is a real country located in Europe. It is the largest country in the European Union and has been the c
ital of France for over 500 years.

>
```

So, what next?

These models are getting smaller, and hardware is getting more capable, we should start to see these capabilities come to “TinyML” scale devices over the next few years.

They work especially well with speech interfaces for products without a screen or keyboard, since they can understand natural language.

Think about boxes on every pillar in a big box store that can tell you where the products you want are, or light switches you can speak to!