**Assignment\_10**

1. What does a SavedModel contain? How do you inspect its content?

**Ans: A SavedModel is a directory containing serialized signatures and the state needed to run them, including variable values and vocabularies**.

1. When should you use TF Serving? What are its main features? What are some tools you can use to deploy it?

Ans: TensorFlow Serving is a flexible, high-performance serving system for **machine learning models, designed for production environments**. TensorFlow Serving makes it easy to deploy new algorithms and experiments, while keeping the same server architecture and APIs.

1. How do you deploy a model across multiple TF Serving instances?

**Ans:**

**Step 1: Install the Docker App.**

**Step 2: Pull the TensorFlow Serving Image. ...**

**Step 3: Create and Train the Model. ...**

**Step 4: Save the Model. ...**

**Step 5: Serving the model using Tensorflow Serving. ...**

**Step 6: Make a REST request the model to predict.**

1. When should you use the gRPC API rather than the REST API to query a model served by TF Serving?

Ans: **When to use gRPC**

1. **Real-time communication services where you deal with streaming calls.**
2. **When efficient communication is a goal.**
3. **In multi-language environments.**
4. **For internal APIs where you don't have to force technology choices on clients.**
5. **New builds as part of transforming the existing RPC API might not be worth it.**
6. What are the different ways TFLite reduces a model’s size to make it run on a mobile or embedded device?

Ans:

**How to compress your Keras model x5 smaller with TensorFlow model optimization**

1. **Train Keras model to reach an acceptable accuracy as always.**
2. **Make Keras layers or model ready to be pruned.**
3. **Create a pruning schedule and train the model for more epochs.**
4. **Export the pruned model by striping pruning wrappers from the model.**
5. What is quantization-aware training, and why would you need it?

**Ans:  Quantization aware training emulates inference-time quantization, creating a model that downstream tools will use to produce actually quantized models. The quantized models use lower-precision (e.g. 8-bit instead of 32-bit float), leading to benefits during deployment**.

1. What are model parallelism and data parallelism? Why is the latter generally recommended?

**Ans: Data parallelism is when you use the same model for every thread, but feed it with different parts of the data; model parallelism is when you use the same data for every thread, but split the model among threads**.

1. When training a model across multiple servers, what distribution strategies can you use? How do you choose which one to use?

Ans: