DL ASSIGNMENT 10

Q1) saved model contains program, trained parameters and computation. Show command allows us to inspect its contents.

Q2) to have same server architecture, tf serving is used. Main feature is integration with tensor flow models.

Q4) gRPC is faster and more rule based.

Q5) tf lite reduces sizes of datatypes like float, int etc.

Q6) training on lower memory occupying datatypes without compromising on accuracy is quantisation aware training. We need it to run tensor flow models on its devices like mobiles.

Q7) in data parallelism, same model is used on different parts of data while in model parallelism, same data is used with all threads where model is split among threads.

Q8) we can use high level APIs like model.fit or custom training loops depending on our requirements.