PER Partners
Challoner House
19 Clerkenwell Close
London EC10RR
London: +44 777 62 555 47
Cairo: +201010954111

MLOps Engineer

22nd Sep 2022

Intro

Despite not being around for long, deep Learning models are now the standard in many fields, from Search Engines to Autonomous Vehicles. DL made their way to the Healthcare sector. At PER partners, we are developing cutting edge AI technology to optimize the operations of the healthcare selector across hospitals, insurance and medical service providers.

Coding Problem:

At PER partners we work with large deep learning models, especially transformers which exceed 100s of millions of parameters. We train, test and deploy those models using different cloud services like AWS or Azure.

In this exercise, you will need to deploy a Hugging Face Transformer-based model. Within Hugging Face,The choice of the model to use is yours. We provided some guidelines below Should you rather go for a non Hugging Face Model, please specify the model and the logic behind the choice.

Technical Notes:

- For implementation use python 3.0+ and their compatible libraries
- This assignment is totally open for the tools/libraries you will use, but please be aware of every tool/library you will end up using and specify why you used it.
- DON'T retrain the model or finetune it, use a pretrained model and deploy it directly for inference.

- Decide on the scoring metric suitable for your model. Feel free to test the model on a large or a small dataset of your choice and provide the details of

- **Model suggestions:**
- AutoModelForSequenceClassification
- Distilbert-base-uncased
- RoBerta
- Distilbert-base-uncased-emotion

Technical Questions:

Question (1):

You have access to an infinite amount of GPUs that have 8 GBs of Memory each to train a deep learning model on.. You try training your model on a batch of size 32, you get a GPU OOM error (out of memory error). You keep decreasing the batch size. To avoid going out of memory, the batch size is reduced to a single image.

a) How many GPUs will you need to train your model for a batch size of 32 ?

b) What is the specific PyTorch or TensorFlow module you will use to conduct your training on? "Write a maximum of 280 characters about how you will use it."

c) You connect the required GPUs, and try to train your model on them. You face the exact OOM error again when you increase the batch size which is 1.. When you check the code, you realize the code has access to only one GPU. You check the hardware connection and it is ok, what do you check next and the associated steps? **Hint: It's probably a software issue outside your deep learning library.**
"Write a maximum of 280 characters elaborating on what you check next."

d)* You started distributed parallel training. You notice the validation score differs significantly between different GPU instances. What do you think could be the reason for this? And what are the steps to address this issue.

Question (2):

Your ML colleagues were training a deep learning model with three versions: small with 70M parameters , distill-model with 100M parameters and large (original) model with 800M parameters. Concurrently, there is a 700GB size of images dataset to train this model on from scratch. The ML team started to train the model on **AWS** using a sagemaker instance, the following issues happened:

1. When the data was loaded, the loader displayed a " memory exceeding' error.. What are the solutions to overcome this error ?
2. Suppose the images dataset is loaded successfully, when loading the model on the GPU, it also displayed an out of memory error but this time, for the GPU. What are the solutions to overcome this error ?
3. Propose the best AWS sagemaker instances to train each of the three versions of the model in the link below
   Hint: https://docs.aws.amazon.com/sagemaker/latest/dg/nbi.html

For inquiries:

- **Ahmed Fakhry:** ahmed.fakhry@perpartners.com
- **Abdelrahman Sharaf:** abdelrahman.sharaf@perpartners.com