Deep Learning on Kubeflow

Praktikum SoSe 2019: Big Data

Iurii Mozzhurin & Rebekka Pech

26.07.2019

1 Introduction

Kubeflow is an open source platform which is based on Kubernetes and was developed by Google. It's dedicated for making deployments of Machine Learning workflows on Kubernetes simple, portable (it works on any Kubernetes cluster) and scalable (upon need). Kubeflow makes it possible to bring ML in any existing cluster.

When talking about Kubeflow you also have to talk about Kubernetes. Kubernetes is an open-source container-orchestration system. With Kubernetes you only need to define what resources the app needs, how many instances should be made, and Kubernetes takes care of deploying, managing and balancing containers on the hardware (nodes).

The main idea of Kubeflow is that it's a ready-to-use set of Machine Learning tools. It's not necessary to install and connect the tools separately, but it can be done automatically. It also provides GUI, which makes the work with them easy for non-programmers.
Kubeflow should be used for distributed ML workflow, for ML in production or for ML in cloud.

2 Use Case

The intention of our use case was to train a CNN model to distinguish between works of ten painters. In the end it should give a (good) prediction whose painting was given in the model.
For this model we used a Dataset from *kaggle* which has ... MB of datas. Because of the size we resized the images and decreased the dataset to ... MB.

For our Use Case we had the following objectives:

1. Creation of a model for transfer learning in Python
2. Creation of a Docker image for this model
3. Training of the model on GKE using TensorFlowJobs on CPUs
4. Training of the model on GKE using TensorFlowJobs on GPUs
5. Distributed training on multiple pods
6. Saving the trained model into a GCS bucket
7. Serving the model with TensorFlowServing

3 Implementation of the Use Case

3.1 Create the model

3.2 Create a Docker

3.3 Training

3.4 Serving

3.5 Results

3.6 Problems

One of our Problems was that we used Keras, but Keras cannot save models/logs directly to the bucket. We had to save it locally and then send it to the bucket.
Also there is no obvious support for distributed learning on one computer we could use.
Another problem was using Kustomize and Ksonnet because they didn't worked.

4 Conclusion

Kubeflow is a tool that is not for data science. It is for model training. If you want to use Kubeflow you have to understand or learn a lot of technologies like Kubernetes, Docker, Keras etc.
Also Kubeflow is still pretty raw - there are still many changes made.
In conclusion we can say that it might be a helpful tool in the future when things are not this raw anymore.