Technology Review

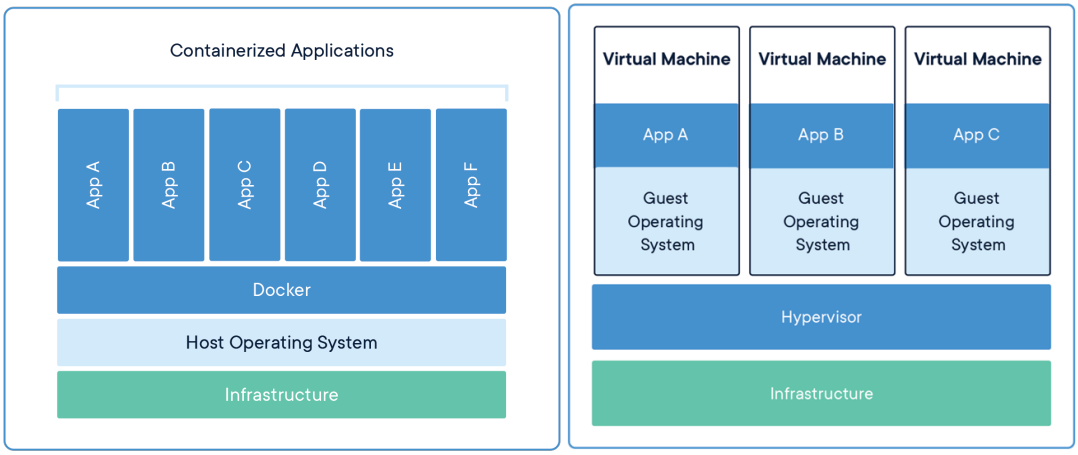
Popular container platforms and their applications in text mining tasks

--- Hecheng Sun

I. Introduction

Containers have gain popularity among the computing community in recent years. They have varied application in many different fields of computer science and provide convenience for computer scientists and engineers in tasks ranging from personal project to large-scale products. In this technology review, we will focus on their applications in machine learning, and more specifically, in text mining tasks.

First, what is a container? According to a good introduction page [1] from Docker, it is a standardize unit of software which packages up code and all dependencies so that we can run the same software on a new computer without worrying the environment setup. A container includes everything you need to run a software. Similar to a virtual machine, it isolates the software you are trying to run from everything else in the environment. However, unlike virtual machines, containers virtualize operating system instead of the underlying hardware, so that they are more portable and efficient. The portability and stability nature of containers make them an ideal way to deploy software from one computer to another with ease, even in large scale.



Illustrations of containers and virtual machine from docker.com [1]

Next, let's look at the most popular container platform right now, Docker, in more details.

II. Most popular container platform: Docker

Docker has a market share of 99% in 2017 and 83% in 2018 [2], making it the most popular container platform in the world. It was launched in 2013 as an open source Docker Engine. It is one of the first to implement the concept of containers and gains popularity quickly in the Linux community. The company behind docker later partnered with Microsoft to create a Windows version of docker for windows servers.

The core idea behind docker is a container image. It is a single and portable file generated from the source computer and can be copied to any other computer that can run Docker Engine. Once you have the image file on the new computer, just run it using Docker Engine, and you will have the same containerized environment, code, libraries, settings from the source computer to run the same software out of the box.

One benefit of using Docker instead of other platform is its size of community. Because it's the most popular container platform, you can find a lot of help from Stackoverflow or other community websites. Most existing container images and files are also in docker as well, so you can download them to your local computer and try them yourself. There is also a core service called Docker Hub [3] which is very similar to GitHub but instead of containing code or other files, Docker Hub's main functionality is to store and share open source docker files to others.

III. Other container platforms

Besides Docker, there are many other container platforms have similar features and functionalities. As introduced in containerjournal.com, five alternatives of docker include CoreOS rkt, Mesos, lxc, OpenVZ, and containerd. All of these five gain more popularities in recent years as developers testing different platforms in their projects. Each of them has some technical advantages comparing to docker, such as a more refined containerization methods for specific environments. As many people in the field expected, even though docker will still be the number one container platform in the next few years, its market share will continue to shrink as more and more alternatives popping up every few months and each of the alternatives focuses on different aspects of containerization to improve upon docker's implementation.

IV. Containers' application in text mining task

Containers like docker provides a great opportunity in the machine learning community. Namely, their portability helps developer to be able to set up training and testing dependencies in a matter of minutes. For the most cases, the user just needs to run the docker image or docker file on the new computer to get all the necessary environment setup correctly. Moreover, according to some extensive testing from numerous users, the performance impact of training in docker is so minimum that it can be ignored in most of the time.

For example, one developer writes a text mining training script on his laptop and wants to deploy it to company's server to perform training. Traditionally, he will need to get a list of all dependency that his script is built on and manually install all packages and libraries with the correct version on the server. The entire process is painfully long and easy to make mistakes. However, with docker, he can just create a docker image from his laptop, then upload and run it on the server. All scripts, data, and dependencies will theoretically be the same by just using one line of command.

Of course, there are times that a simple docker image does not work perfectly on the new machine, which is why an alternative is preferred by most developer, namely, dockerfile [4]. A dockerfile is a text document that contains a series of commands to build a docker image. Each

line of the dockerfile is very similar to a Linux command that you normally use to install the package you want in command line. The downside is that developers will need to compose this kind of dockerfiles manually, but the advantage is it will be much smaller than a docker image and should have better compatibility on the new computer environment.

V. Conclusion

Container platforms like docker are essential in machine learning and text mining community. It provides excellent portability to help developers to deploy their code and environment on a new machine with ease. As the development of container grows rapidly, people can expect containers to play an even larger role in ML/AI in the near future.

VI. Reference

[1] https://www.docker.com/resources/what-container
[2] https://containerjournal.com/topics/container-ecosystems/5-container-alternatives-to-docker/
[3] https://hub.docker.com
[4] https://docs.docker.com/engine/reference/builder/