DEEP LEARNING IN PRODUCTION

Sergios Karagiannakos

AI SUMMER

DEEP LEARNING IN PRODUCTION

Sergios Karagiannakos

AI SUMMER

Table of Contents

Preface

Deep Learning in Production is a product of one year of effort. The pages and the code you will read, begun as articles on our blog AI Summer and they were later combined and organized into a single resource. Some were rewritten from scratch; some were modified to fit the book's structure and some they are entirely new.

The reason I decided to invest the time in writing this book is very simple. The practices and principles mentioned here is what I wish I knew when I started my journey on machine learning. Having an all-in guide that outlines every aspect of the deep learning pipeline would have accelerate my learning curve by a big margin. I do hope that it will do the same for you. Or at least, give you an overview of all the different components and steps, so you can have a holistic view of the field.

It accumulates the knowledge I have gained over the past years by working both as part of the machine learning infrastructure team at HubSpot, as a Data Scientist in a web agency, and as an independent contractor with various start-ups. Each project helped me learn something new and each team gave me a fresh and unique perspective on the field. This is what I hope to transmit to you.

The background and skills that you will acquire from this book will provide you better job opportunities, will differentiate you from other data scientists and machine learning researchers, and the most important thing: they will make you a better and more well-rounded engineer.

Thank you for choosing this book. I deeply appreciate it.

Sergios Karagiannakos

Chapter 1

About this Book

In this chapter:

- What this book is about
- What do you need to know beforehand
- Who should read this book

1.1 Welcome to Deep Learning in Production

Hello. It's amazing to have you here. I'm super excited and you should be too. Deep Learning is one of the most transformative technologies of today. Deep neural networks are gradually changing many industries from healthcare and transportation to manufacturing and agriculture. However, the gap between creating a novel neural network architecture and using the model in an actual product is still quite large. The deployment of a machine learning model into production is where the fun begins. This is what this book is all about. **We will discuss the basic steps one should take to build, deploy, scale, and maintain deep learning models**.

1.2 Is this book for me?

Since you are here, I will probably say yes. But let's break it down. If you are interested in learning how to take a simple model and transform it into a real-world

application, then the answer is absolutely yes.

Software engineers

If you come from a software engineering background and you are starting out with deep learning, you will learn how to apply the things you already know on AI projects.

Machine Learning researchers

If you don't have a programming background and you are a deep learning researcher or practitioner, you will familiarize yourself with the software development lifecycle. You will learn the basic principles of how to deploy and scale a machine learning application. In the end, you will be able to create an entirely new deep learning app from scratch and serve it to actual users.

Machine Learning engineers

If you are already a machine/deep learning engineer, you will solidify your understanding of the field, you will explore different practices and ideas, and you will familiarize yourself with software skills and technologies that you might have missed.

Data Scientists

If you are an experienced or aspiring Data Scientist, you will understand what it takes to productionize your models and to build customer-facing applications. You will comprehend the technical debt machine learning has, and you will gain a much broader image of the field.

1.3 What is the book's goal?

If I were to summarize the book into 4 sentences, I'd say that you will learn:

- how to structure and develop production-ready machine learning code
- how to optimize the model's performance and memory requirements
- how to make it available to the public by setting up a service on the cloud
- how to scale and maintain the service as the user base grows

1.4 Will this be difficult to learn?

To be honest with you, it won't be very easy. We cover some advanced concepts and technologies in this book, which might be hard to understand at first. We introduce

a variety of frameworks and libraries that may have a steep learning curve. But don't be discouraged by that. As a machine learning engineer, no one expects you to be an expert on DevOps. However, having a basic understanding of all the systems and concepts, will get you a long way.

1.5 Why should you read this book?

The reason is very simple: because deep learning models are useless if they aren't applied in real-life applications. Don't get me wrong, research is awesome. No one enjoys creating a new fancy, state-of-the-art-model more than me. But the ultimate goal is usually to make the model visible to real users. And that's exactly what you'll learn here.

To clarify why Software Engineering is an undeniable significant part in deep learning, take for example Google Assistant. Behind Google Assistant is, without a doubt, an ingenious machine learning algorithm. But do you think that this amazing research alone is capable of answering the queries of millions of users at the same time? Absolutely not. There are dozens of software engineers behind the scenes, who maintain, optimize, and build the system. This is exactly what we are about to discover.

1.6 How to use this book?

This isn't going to be one of those high-level, abstract books that talk too much without practical value. Here we are going to dive deep into software, we will analyze details that may seem too low level, we will write a lot of code, and we will present the full deep learning development cycle from start to end.

Note, though, that all of the practices mentioned in this book are merely suggestions that come from my experience. Nothing will ever be declared as "the best way to do things".

There are many best practices that, in my opinion, will make your job much easier but that's up to each engineer to decide. After all, this is what makes our job so much fun. The ability to design and build things just the way we want.

Always remember that machine learning systems are not very mature yet, and new solutions and frameworks are created at a rapid pace.

1.7 How is the book structured?

Each chapter of the book will focus on a particular phase of the ML lifecycle.

We will:

1. Start from designing our end system.
2. Explore best practices on how to develop and structure deep learning code.
3. Dive more into the data pre-processing step.
4. Learn how to build optimized data pipelines.
5. Learn how to train models locally or in the cloud.
6. Build a first version of our deployable service.
7. Deploy it in a cloud instance.
8. Discover how to scale and maintain the system.
9. Close with existing end-to-end solutions and efforts to standardize the entire lifecycle.

It's going to be a long but amazing journey.

1.8 Do I need to know anything else before I get started?

Needless to say, that in order to follow along, you will need to have a solid background on programming and/or machine learning.

We will use Tensorflow 2.0 in order to showcase the different techniques and ideas. So, a good understanding of Tensorflow would be ideal. However, all the principles apply to other frameworks such as Pytorch, MXNet and CNTK.

Besides Tensorflow, here is a list of other frameworks and tools that will use throughout the book: Flask, uWSGI, Nginx, Docker, Kubernetes, Google Cloud. As you have probably guessed, this book is quite opinionated in terms of libraries. You can find links with the documentation of all mentioned libraries in the Appendix.

But make no mistake, the choice of the particular tools is only to showcase the concepts. In your own projects, you are of course free to use whatever you like. Along the process, I will also give you some extra tips to better utilize the tools, increase your productivity and enhance your workflow.

Without further ado, let's get started.

Index

About the Author

Sergios Karagiannakos is a Machine Learning Engineer with a focus on ML infrastructure and MLOps.

He is based in Athens, Greece and graduated with a Master's in Electrical and Computer Engineering from University of Patras. He has helped several companies build their Artificial Intelligence products; most notably, Eworx SA where he joined as a Data Scientist, and Hubspot as part of the Machine Learning infrastructure team.

In 2019, he founded AI Summer, an educational platform around Deep Learning. During this time, he has authored more than 50 articles and he published the Introduction to Deep Learning & Neural Networks course.

Interesting facts: He was included in the Top 100 influential voices and brands in Data Science and Deep Learning, he strives to bring the entire Greek tech community together, and he really wishes that Artificial General Intelligence will be solved in our lifetime.