Chapter 11: Generative AI in GCP

In the preceding two chapters we have demonstrated AWS and Azure capabilities within the realm of generative AI; we now turn our attention to Google offering in that space. What differentiates the company from its main competitors in the space – Amazon and Microsoft – is that it is currently the only one offering a comprehensive suite of commercial solutions covering every aspect of the AI-focused suite of tools. From dedicated AI chips, through computing power, to development tools and applications – Google has a complete AI stack that serves as a foundation for its offering. Their proprietary Tensor Processing Units (TPU) are designed specifically for machine learning tasks, optimized for use with TensorFlow, and are boasting faster and more efficient performance for AI workloads; while GPU (and CPU) units have not been dethroned across the industry, over the last few years TPU have emerged as formidable rivals.

The Cambrian explosion of generative AI started in the second half of 2022 and the initial impression was that Google had to play catch-up to Microsoft / OpenAI (especially in the NLP space, due to chatGPT parabolic rise). Upon closer examination, it seems like Google strategy has been more of a comprehensive enterprise-grade approach. We begin by discussing the educational offer on GCP: if you, dear reader, have reached this point in the book, chances are you are already familiar with the basics; nevertheless, the series of mini courses offered on the platform can be useful complement to your knowledge.

Having reviewed the educational offer, we will examine the capabilities of Generative AI Studio. The application is built on the Vertex AI platform and allows developers to create their own generative AI apps in text and vision domains. More advanced users can make us of the Model Garden, which gives access to a collection of pretrained models, for advanced exploration and interaction.

If you prefer to hear the summary of Google Generative AI offerings from the company themselves, feel free to check out the introductory videos: <https://www.youtube.com/watch?v=YCZ6nwGnL4o>. Once you have done that, you can come back here and we will embark on the next stage of our journey.

# Learning path

The educational provisions on GCP are designed for a wide range of individuals: from novice enthusiasts eager to expand their understanding to seasoned professionals seeking to refine their expertise. These crash courses allow for an a la carte approach, allowing learners to delve into specific topics – following them sequentially can be helpful in certain cases, but is not required. The first step is to go to the training resources page: <https://cloud.google.com/blog/topics/training-certifications/new-google-cloud-generative-ai-training-resources>

You will be greeted by the screen shown in Figure 11.1 below. Scroll down to “Generative AI Learning Path” and click on the link. This will redirect you to <https://www.cloudskillsboost.google/journeys/118>, where you begin your education.

A picture containing text, screenshot, diagram, design

Description automatically generated

Figure 11.1: Landing page for the Generative AI track from GCP

# A screenshot of a google cloud Description automatically generated with low confidence

Figure 11.2: Generative AI Learning Path page

Each micro-course is composed of three parts: video content, reading list and a quiz to test your newly acquired knowledge. We briefly summarize the modules below:

1. **Introduction to Generative AI**: this course defines Generative AI and explains how it is different from other domains of ML. In addition, it covers major types of models used in the field and describes popular AI applications. Even if you are an experienced ML practitioner, spending half an hour on refreshing the basic definitions is a good idea.  
     
   An important disclaimer needs to be made at this point: *while the structure of the courses is likely to remain stable, the content itself might have changed a bit by the time you are reading these words (e.g. which models are used are integrated, or specific details of the interface); the reason for that is the incredibly fast pace of change in the field: nary a fortnight seems to go by without somebody releasing an ever-more-powerful LLM, beating a previous SOTA. “Change is the only constant” is a rather terrible cliché, but in the context of generative AI it is quite appropriate*.
2. **Introduction to Large Language Models**: this course delves into the concept of LLM, their various applications and utilization of prompt tuning to improve LLM performance. In addition, Google tools useful for personalized Gen AI applications are reviewed. This one is notable for its scope: starting from LLM types (generic, dialog-tuned, instruction-tuned), it discusses also advanced concepts like Chain of Thought.
3. **Introduction to Responsible AI**: This introductory microlearning course is aimed at explaining what responsible AI is, why it's important, and how Google implements responsible AI in their products. It also introduces Google's 7 AI principles.
4. **Introduction to Image Generation:** if population averages are anything to go by, chances are you, dear reader, got into generative AI because you saw an image created with Dall-E or somesuch – which means you have a pretty good idea of what they can do. In this course you will get an introduction into how they do it: diffusion models draw inspiration from thermodynamics and over the last few years they have gained enormous popularity both in research and in the industry. You will get a chance to dive into the theory underpinning diffusion models, as well as learn how to effectively train and deploy them on Vertex AI.
5. **Encoder-Decoder Architecture:** This course provides an overview of the encoder-decoder architecture, a widely used and effective machine learning framework for tasks involving sequences, such as machine translation, text summarization, and question answering. You will gain knowledge about the key elements of the encoder-decoder architecture, including training and deploying these models. During the accompanying lab tutorial, you will code a basic implementation of the encoder-decoder architecture in TensorFlow for generating poetry from scratch.
6. **Attention Mechanism**: in this module you will became acquainted (or re-acquainted – the original paper came out in 2017, after all) with the attention mechanism: a powerful technique enabling DL algorithms to concentrate on segments of an input sequence. You will gain an understanding of how attention operations and how it is applied in enhancing the efficiency of various ML tasks.
7. **Transformer Models and BERT Model:** the course introduces you to \*the\* application of attention mechanism – Transformer architecture, along with its most famous example: Bidirectional Encoder Representations from Transformers (BERT) model. You will learn about the main components (self-attention), and how they comprise the BERT together.
8. **Create Image Captioning Models:** In this course, you will learn to build an image captioning model using deep learning techniques. You will gain knowledge about the various elements comprising an image captioning model, including the encoder and decoder, as well as the process of training and evaluating your model. By the conclusion of the course, you will have the ability to develop your own image captioning models and utilize them for generating descriptive captions for images.
9. **Introduction to Generative AI Studio:** Generative AI Studio is a very interesting offering from GCP, so this final course serves as an excellent segue into our next section.

# Generative AI Studio

Generative AI Studio is a GC console tool for rapidly prototyping and testing generative AI models. You can test sample prompts, design your own prompts, and customize foundation models to handle tasks that meet your application’s needs.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 11.3: Generative AI Studio landing page

Asdadsadasdasd

Asdasd

A screenshot of a web page

Description automatically generated with medium confidence

Figure 11.4: XX

# Generative AI App Builder

Asdasd

# Model Garden

<https://www.youtube.com/watch?v=I7UiSU96CLc>

aasd

# Running models on GCP

Asd