



Slide

深入淺出 Google Gemma LLM 模型

Simon Liu

Google Cloud Summit Taipei 2024

2024/06/13



About me: 劉育維 / Simon Liu



- Ocard - AI 工程師
- AI, ML, DL, LLM Architect and Engineer / Technical Writer / Speaker
- Try to use AI to do the application and help to solve the pain point by AI methods.
- My Personal Information:



Linkedin



Personal Website



Today's Topic

- 1 Gemma 模型介紹
- 2 如何使用 Gemma 模型
- 3 透過 sentence transformer 和 Google Gemma 模型, 進行 RAG 應用
- 4 結語



技術隨時在變化

請依照工具提供的官方資訊為主, 我已盡力做好所有查核處理工作。

Part 1

Gemma 模型介紹

先問問看，大家是否玩過 **Gemma LLM** 模型呢？



Gemma LLM 模型的自我介紹



Medium 介紹

- 第一次釋出日期: Feb 21, 2024
- 模型名稱: Gemma 系列模型
 - 此模型由 Google 官方所開源出來的 SOTA AI 模型
- 模型開源狀況 / License:
 - Gemma 目前採用 Google 所撰寫的 License 授權方式 – Gemma Terms of Use, 統整相關內容並理解後, 是一個可商用的模型。

都已經有了 Gemini 模型，為何要釋出 Gemma 模型？

- 社群回饋機制
- 邊緣 LLM 服務應用可能性
- 可供研究人員和開發人員用於各種目的，包括：
 - a. 探索新的 LLM 應用
 - b. 開發新的 LLM 技術
 - c. 使 LLM 更易於獲取

Gemma 目前的模型能力比較 (等待 Gemma v2 的更新)

	Meta Llama 3 8B	Meta Llama 2 7B	OpenAI GPT-4	Google Gemma 7B-it	Google Gemini 1.5 Pro	MistralAI Mistral 7B Instruct
Open Source / Close Source	Open Source	Open Source	Close Source	Open Source	Close Source	Open Source
MMLU (5-shot)	68.4	34.1	86.4	53.3	81.9	58.4
GPQA (0-shot)	34.2	21.7	35.7 (0-shot CoT)	21.4	41.5 (0-shot CoT)	26.3
HumanEval (0-shot)	62.2	7.9	67.0	30.5	71.9	36.6
GSM-8K (8-shot, CoT)	79.6	25.7	92.0 (5-shot CoT)	30.6	91.7	39.9
MATH (4-shot, CoT)	30.0	3.8	52.9	12.2	58.5	11.0



PaliGemma 介紹

目前 Gemma 種類

Gemma v1 / v1.1 / v2 (coming soon)

- 高效能且輕量化的大型語言模型
- 主要可做對話式大型語言模型
- 各項評估上表現優良

RecurrentGemma

- 支援研究人員進行大批次的高效推理, 採用循環神經網路和局部注意力機制提升記憶效率。
- 基準測試上成績與Gemma 2B模型相當, 但是RecurrentGemma使用的記憶體量更少

CodeGemma

- 針對開發人員和企業的程式碼完成、生成和聊天工具使用情境
- 多程式語言能力, 主要以Python、JavaScript、Java等各種熱門程式語言的程式碼撰寫建議

PaliGemma

- 視覺語言開放模型, 能夠針對影像字幕、視覺問答、理解圖像內文字、物件偵測、物件切割的應用案例提供最佳化

大小模型的分工任務

- 大模型：期待能夠分解任務，並且讓任務能夠被定義清楚。
- 小模型：完成指派的任務，並且任務能夠交付至大任務中。

-> 小型的企業組織就出現了。

Gemini Model

CodeGemma

Gemma

PaliGemma

未來生成式 AI 趨勢 by 簡立峰 (曾擔任Google台灣區分公司總經理等職務)

Tech Trends to Watch

- 模型網路化
 - 大模型化 (Gemini Ultra)
 - 小模型化 (Google Gemma)
 - 從單一到多模生態 (Multi-Modal - PaliGemma)
- 雙螢應用+AI
 - Doc, Excel, PPT, Mail, Photo, Chat, Search, ... "Agent"
 - Application: Google Workspace + Gemini
- Edge AI 新機載
 - AI phone (Google Pixel), ...
- 機器人月球化
 - LLM生數據／人機對話交替
 - 白領到藍領全面影響



Part 2

如何使用 Gemma 模型

你有幾種方式使用 Gemma

套件程式碼撰寫



寫程式碼，讓模型按照套件方式啟動

相關工具啟用



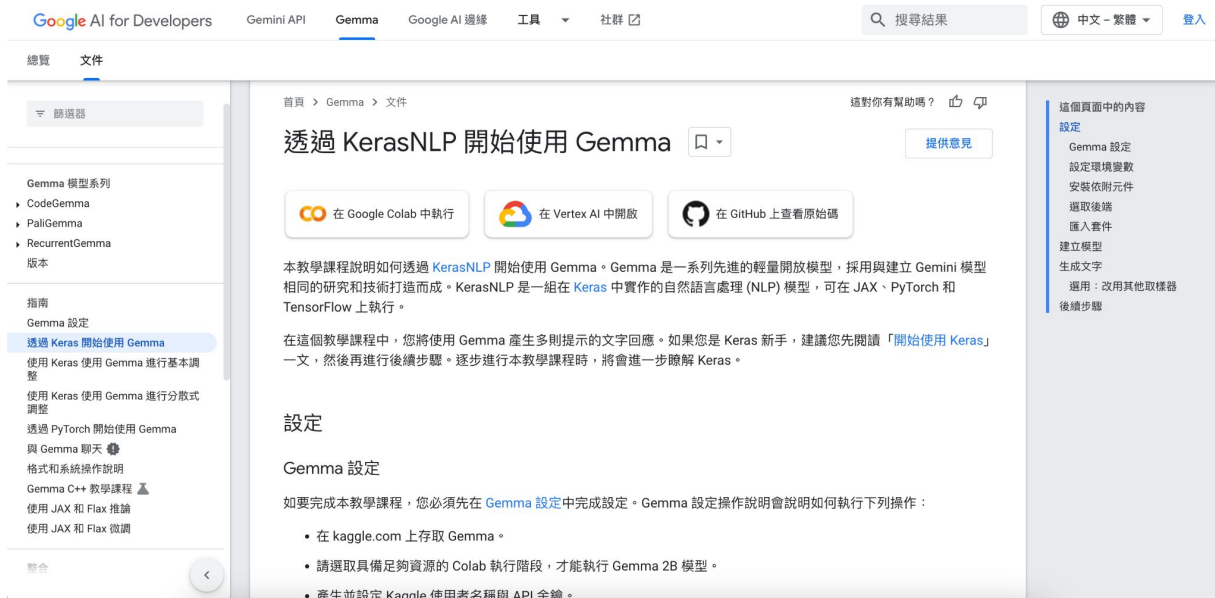
透過工具啟動服務，透過 API 來使用

Gemma Model in KerasNLP

- 專門用於自然語言處理(NLP)的 Keras 擴展庫
- 提供了一系列的工具和模組來簡化和加速NLP 任務的開發。



文件介紹



The screenshot displays the Google AI for Developers website, specifically the 'Gemma' section. The page is titled '透過 KerasNLP 開始使用 Gemma' (Getting started with Gemma using KerasNLP). It features a navigation sidebar on the left with links to 'Gemma 模型系列' (Gemma model series), '指南' (Guides), and 'Gemma 設定' (Gemma settings). The main content area includes a '透過 Keras 開始使用 Gemma' (Getting started with Gemma using Keras) section, which provides a brief overview of the model and its capabilities. Below this, there is a '設定' (Settings) section that outlines the steps to get started with the model, including installing the necessary dependencies and setting up the environment. The page also includes a '提供意見' (Provide feedback) button and a '登入' (Login) link in the top right corner.

Google AI for Developers Gemini API Gemma Google AI 通緣 工具 社群

搜尋結果 中文 - 繁體 登入

總覽 文件

透過 Keras 開始使用 Gemma

在 Google Colab 中執行 在 Vertex AI 中開啟 在 GitHub 上查看原始碼

這個教學課程說明如何透過 KerasNLP 開始使用 Gemma。Gemma 是一系列先進的輕量開放模型，採用與建立 Gemini 模型相同的技術打造而成。KerasNLP 是一組在 Keras 中實作的自然語言處理 (NLP) 模型，可在 JAX、PyTorch 和 TensorFlow 上執行。

在這個教學課程中，您將使用 Gemma 產生多則提示的文字回應。如果您是 Keras 新手，建議您先閱讀「開始使用 Keras」一文，然後再進行後續步驟。逐步進行本教學課程時，將會進一步瞭解 Keras。

設定

Gemma 設定

如要完成本教學課程，您必須先在 Gemma 設定中完成設定。Gemma 設定操作說明會說明如何執行下列操作：

- 在 kaggle.com 上存取 Gemma。
- 請選取具備足夠資源的 Colab 執行階段，才能執行 Gemma 2B 模型。
- 產生並設定 Kaggle 使用者名稱與 API 金鑰。

這個頁面中的內容

- 設定
 - Gemma 設定
 - 設定環境變數
 - 安裝依附元件
 - 獲取後端
 - 匯入套件
- 建立模型
 - 生成文字
 - 選用：改用其他取樣器
- 後續步驟

Ollama

- Ollama 是一個開源軟體，讓使用者可以在自己的硬體上運行、創建和分享大型語言模型服務。
- 這個平台適合在地端運行模型，因為它不僅可以保護隱私，還允許使用者透過命令行介面輕鬆地設置和互動。
- Ollama 支援非常多種模型，並提供彈性的客製化選項，例如從其他格式導入模型並設置參數。



Gemma Model in Ollama

文件介紹



gemma

Gemma is a family of lightweight, state-of-the-art open models built by Google DeepMind. Updated to version 1.1

2B 7B

↓ 1.6M Pulls ⌚ Updated 6 weeks ago

7b	102 Tags	ollama run gemma
Updated 6 weeks ago		a72c7f4d0a15 · 5.0GB
model	arch gemma · parameters 8.5B · quantization Q4_0	5.0GB
license	Gemma Terms of Use Last modified: February 21, 2024 By usi...	8.4kB
template	<start_of_turn>user {{ if .System }}{{ .System }} {{ end }}...	136B
params	{"penalize_newline":false,"repeat_penalty":1,"stop":["<sta...	109B

Gemma is available in both **2b** and **7b** parameter sizes:

- ollama run gemma:2b
- ollama run gemma:7b (default)

README MIT license

Ollama Python Library

The Ollama Python library provides the easiest way to integrate Python 3.8+ projects with [Ollama](#).

Prerequisites

You need to have a local ollama server running to be able to continue. To do this:

- Download: <https://ollama.com/>
- Run an LLM: <https://ollama.com/library>
 - Example: `ollama run llama2`
 - Example: `ollama run llama2:70b`

Then:

```
curl https://ollama.ai/install.sh | sh
ollama serve
```

Next you can go ahead with `ollama-python`.

Install

```
pip install ollama
```

Usage

```
import ollama
response = ollama.chat(model='llama3', messages=[
  {
    'role': 'user',
    'content': 'Why is the sky blue?',
  },
])
print(response['message']['content'])
```


Part 3

透過 Sentence Transformer 或 Google Cloud 服務和 Google Gemma 模型, 進行 RAG 應用

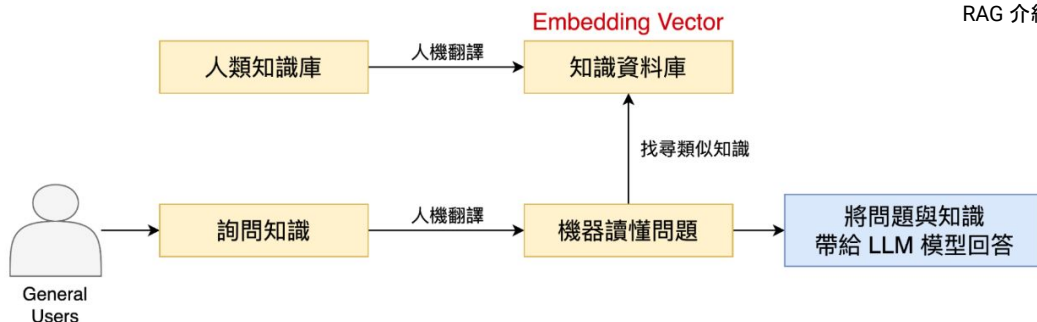
RAG / Fine-Tune



RAG 介紹

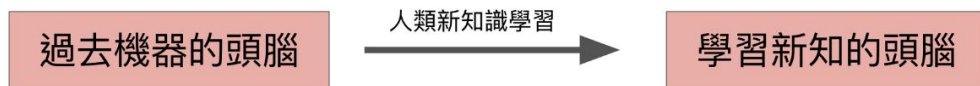
RAG (Embedding)

當人類詢問問題時，找尋說明書了解知識後，再回覆問題。



Fine-Tune Model

類似小孩子學習新知的概念，經過學習，就能夠得到新知。



RAG 和 Fine-Tune 比較

	Fine-Tune 微調模型	RAG (Embedding)
比喻	就像考試前認真讀書，考試 closed book 去回答考試題目。	就像考試 open book，帶筆記去考試，若筆記上有寫可以回答的很好
缺點	訓練模型需要花時間和計算成本，不可能隨時訓練更新資料	仍有 Token 長度限制 要用工具抓資料因此處理時間較長
優點	品質可能更好，這需要機器學習專業知識	不用擔心新資料更新

到底什麼時候 Fine-Tune ?

以法律相關背景做舉例：

法律判例機器人

VS

法律條文機器人

透過過去判決經驗，來判斷
是否可以推論問題內容



根據過往案例來判斷



RAG 讓判例能夠快速準確被
檢索出來，且經常會被更新

透過法律條文，來了解此法
律案件違反哪條規定



所有法律條文綜合判斷



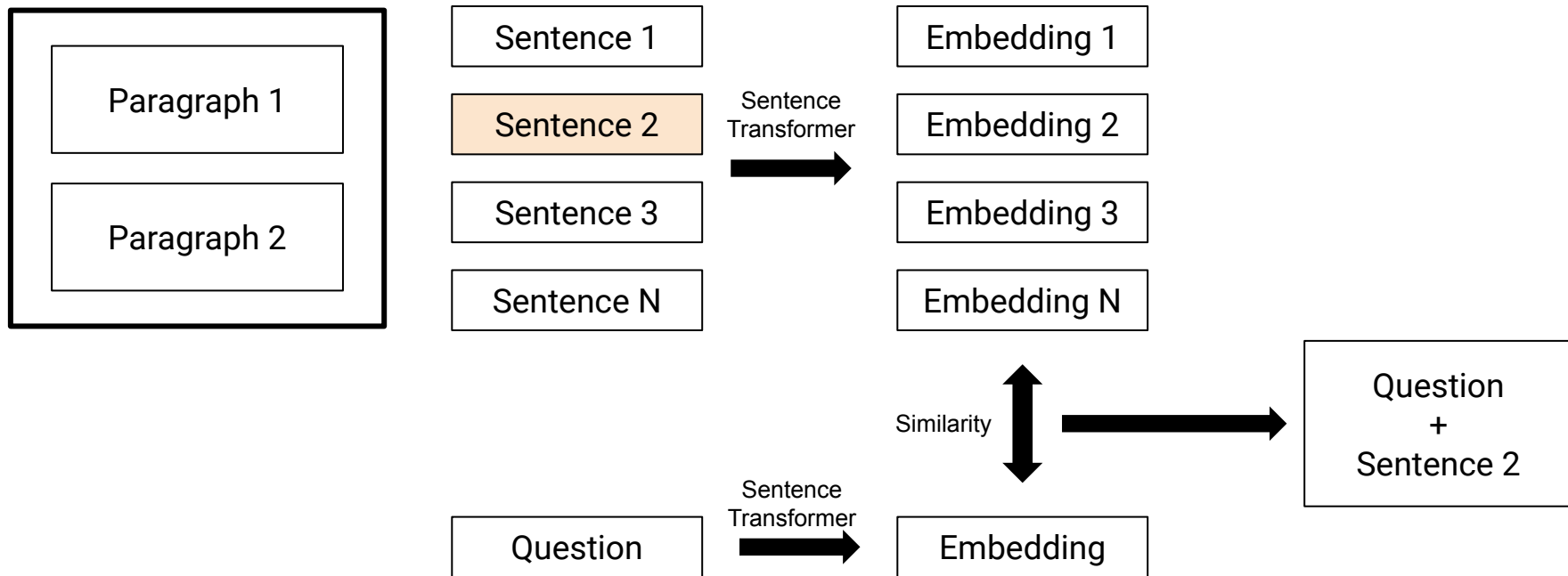
先 Fine-Tune 讀懂條文

Sentence Transformer

專為句子和段落 Embedding 而設計的模型，它能夠以高效的方式計算句子的向量表示。



文章介紹



Google Cloud 也支援 RAG 相關服務

- Google Cloud Vertex AI Vector Search
 - Google Official Doc: [Google Cloud Vertex AI Vector Search](#)
- Bigquery:
 - Google Official Doc: [Bigquery](#)
- Pgvector:
 - Google blog: [Building AI-powered apps on Google Cloud databases using pgvector, LLMs and LangChain](#)

Part 4

結語

Gemma 到底可以用在什麼地方？

- Gemma 目前的能力大概就是國小生能力，能做的事情不多。
- 六月份更新後，預期可以超越 GPT 3.5，接近 GPT 4 能力，大約是一個工讀生能力。
- 觀察一下公司裡面工讀生在做什麼？
 - 分類文件
 - 文件理解
 - 基礎知識問答

結論

- Gemma 在六月份會迎來一次更新, 相信在能力上能夠與其他開源模型一拼
- 透過工具來啟動 Gemma 服務
 - Ollama / Llama-cpp / VLLM 等
 - KerasNLP / Transformer
- 透過 RAG 等方式, 讓產品化能夠做的更好
 - Sentence Transformer
 - 第三方工具
- Fine-Tune?
 - 成本高, 特殊條件下再思考是否做 Fine-Tune 處理。



Linkedin



Personal Website



Slide



Thanks for listening!

Simon Liu
Google Cloud Summit Taipei 2024
2024/06/13