Build
with AI

Gemma简介

Google Developer Groups
天津

内容提要

01 生成式AI简介

02 Gemma简介

03 开始使用Gemma

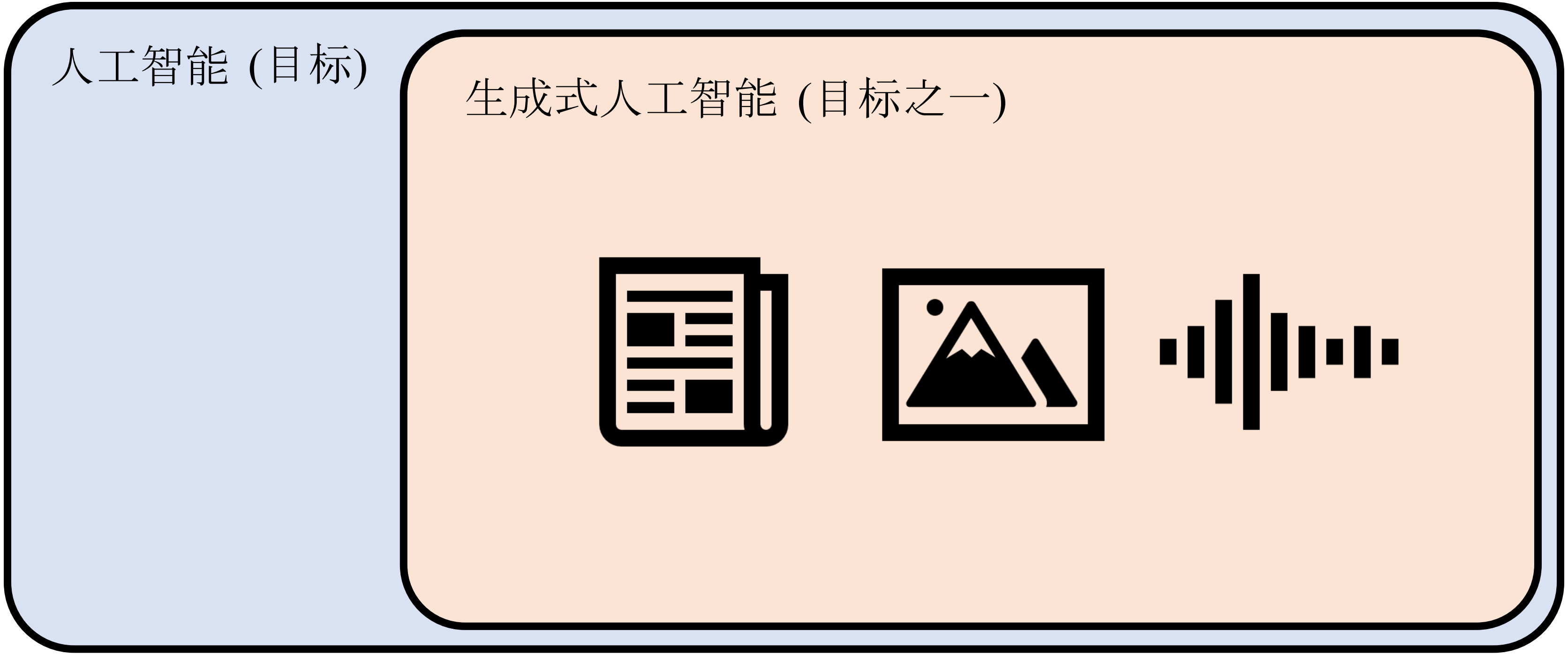
04 Gemma新进展

Build
with AI

Google Developer Groups
天津

生成式AI简介

生成式人工智能

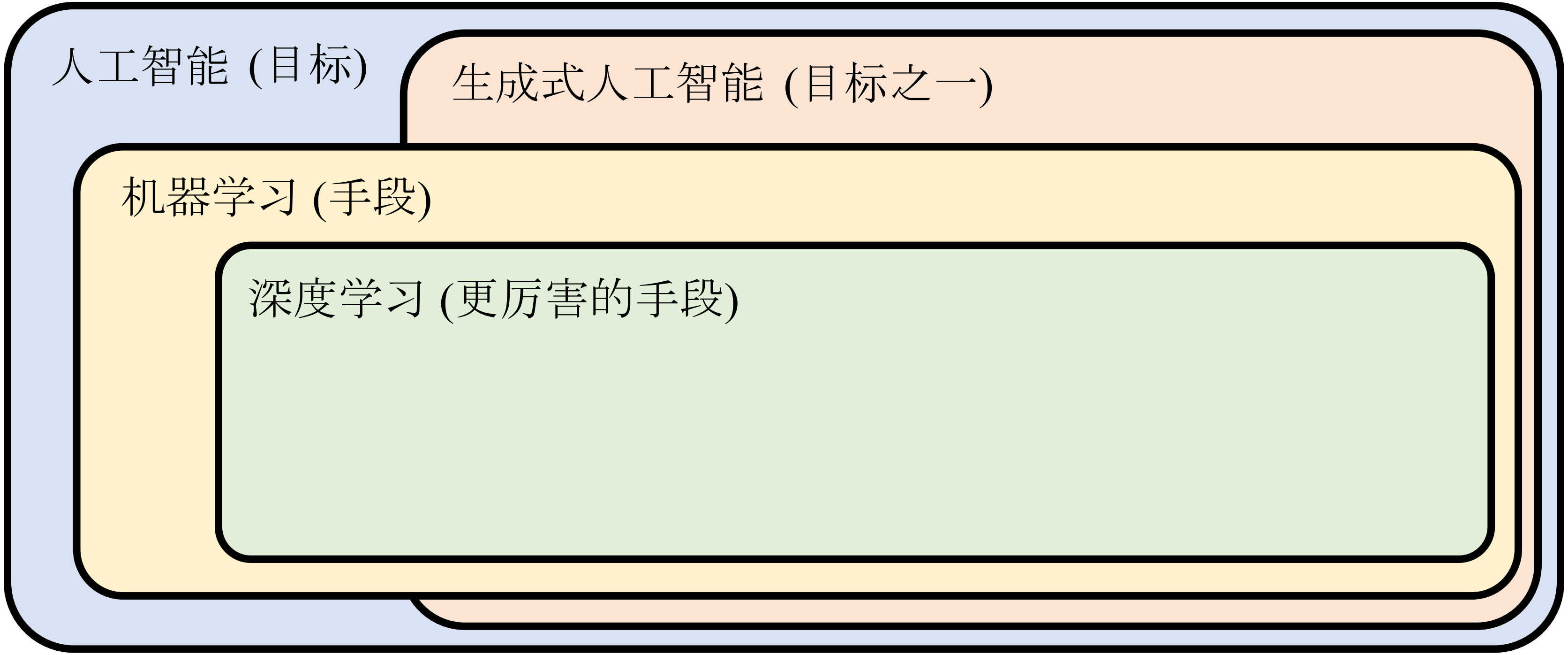


人工智能 (目标)

生成式人工智能 (目标之一)

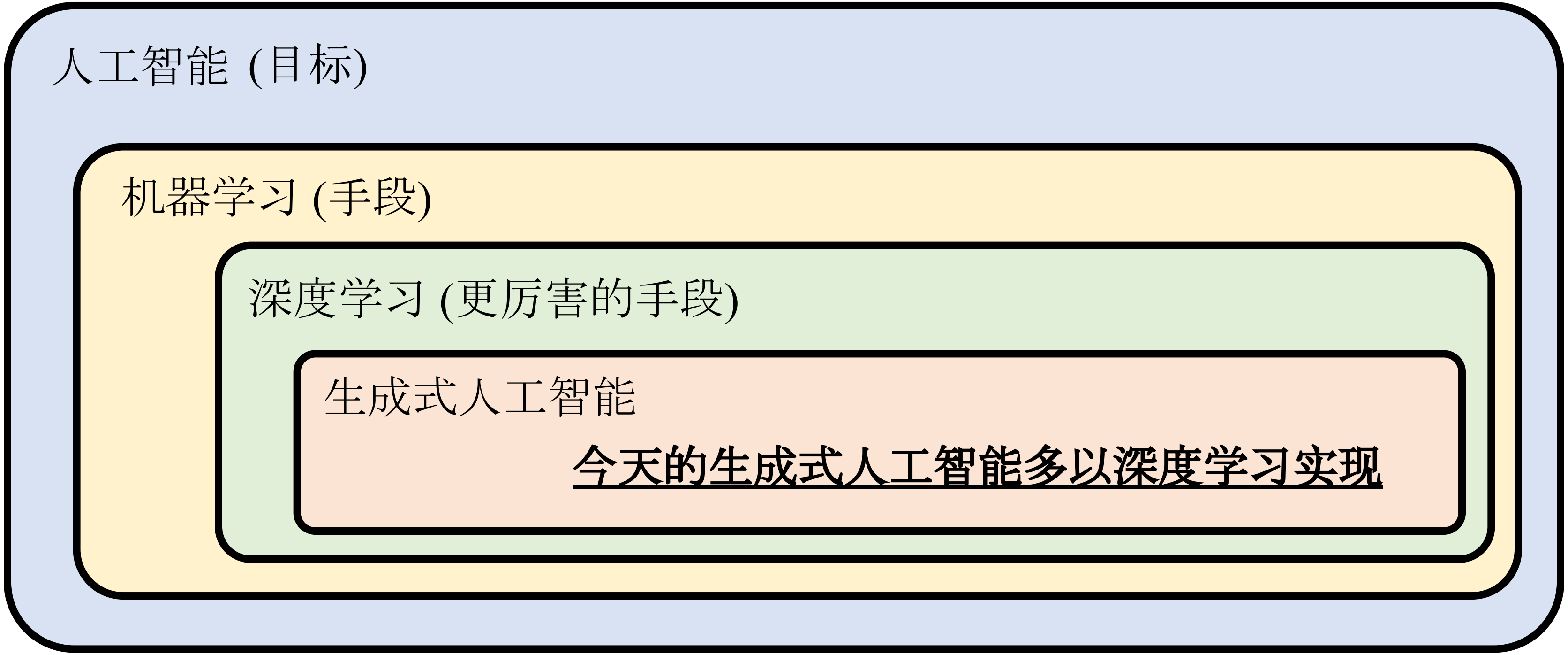
本页Slide来自李宏毅 (Hung-yi Lee)教授2024年《生成式人工智慧导论》第1讲“生成式AI是什么”幻灯片

生成式人工智能



本页Slide来自李宏毅 (Hung-yi Lee)教授2024年《生成式人工智慧导论》第1讲“生成式AI是什么”幻灯片

生成式人工智慧



人工智能 (目标)

机器学习 (手段)

深度学习 (更厉害的手段)

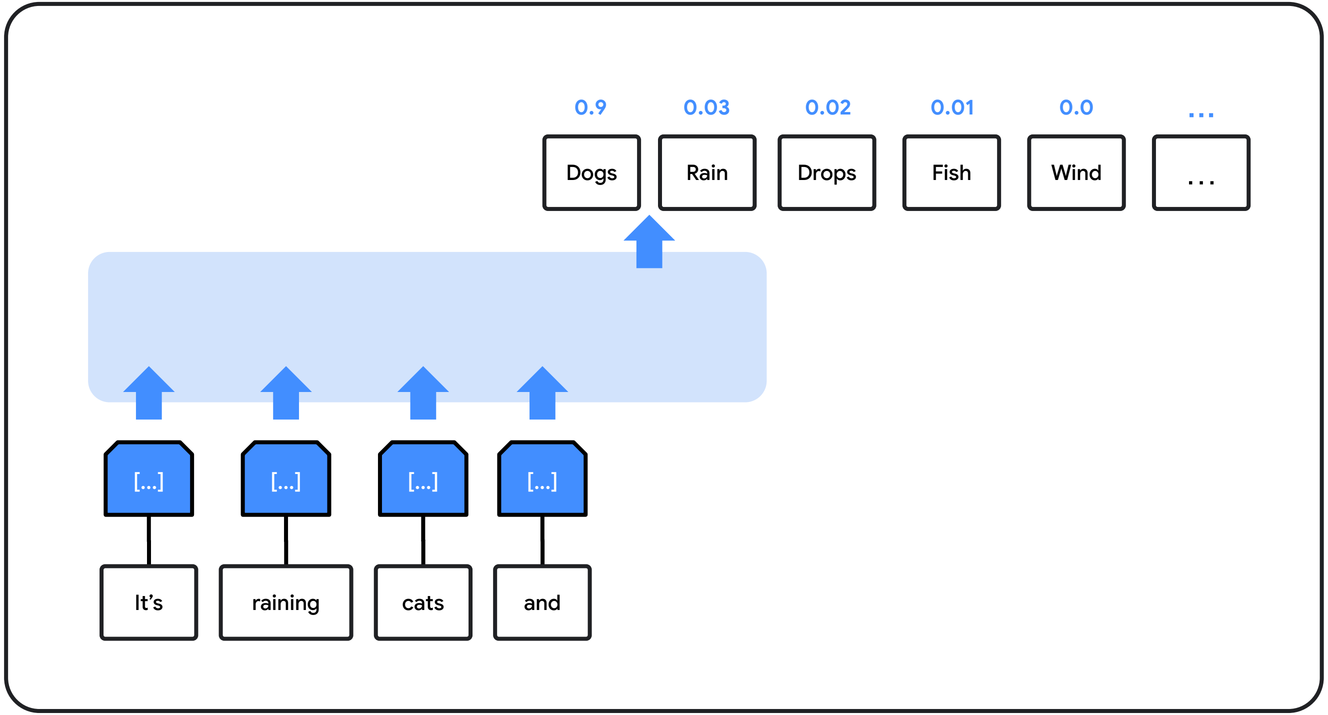
生成式人工智能

今天的生成式人工智能多以深度学习实现

本页Slide来自李宏毅 (Hung-yi Lee)教授2024年《生成式人工智慧导论》第1讲“生成式AI是什么”幻灯片

LLMs Explained

What is an LLM?



Roses are red,

LLMs Explained

Roses are red,
Violets are blue,
Sugar is sweet,

```
for(var i = 0
for(var i = 0; i < 10; i++) {
```

Build
with AI

Google Developer Groups
天津

Gemma简介

模型结构

- **多查询注意力（Multi-Query Attention）**

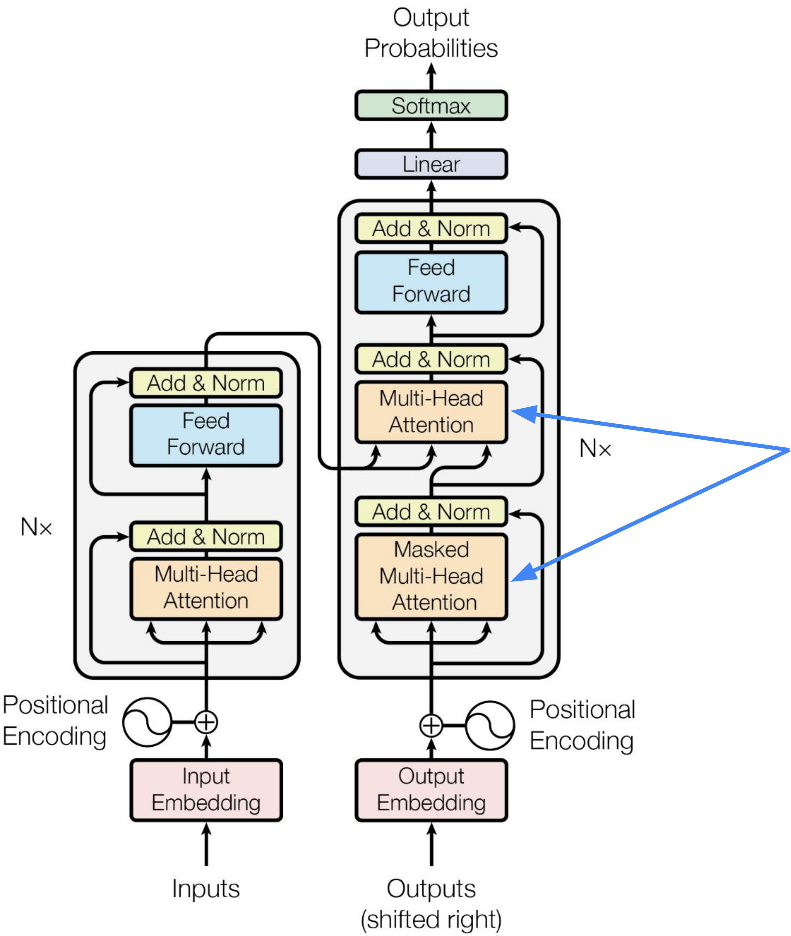
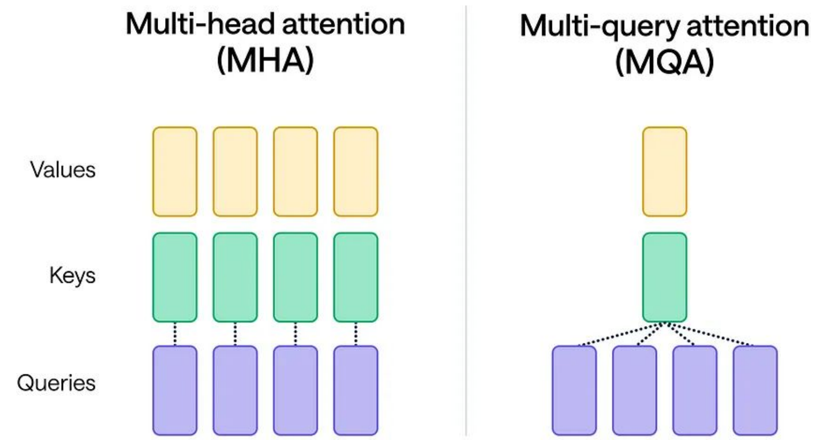


Figure 1: The Transformer - model architecture.



左图来源：[Attention Is All You Need]

右图来源：[Grouped Query Attention (GQA) vs. Multi Head Attention (MHA): Optimizing LLM Inference Serving]

模型结构

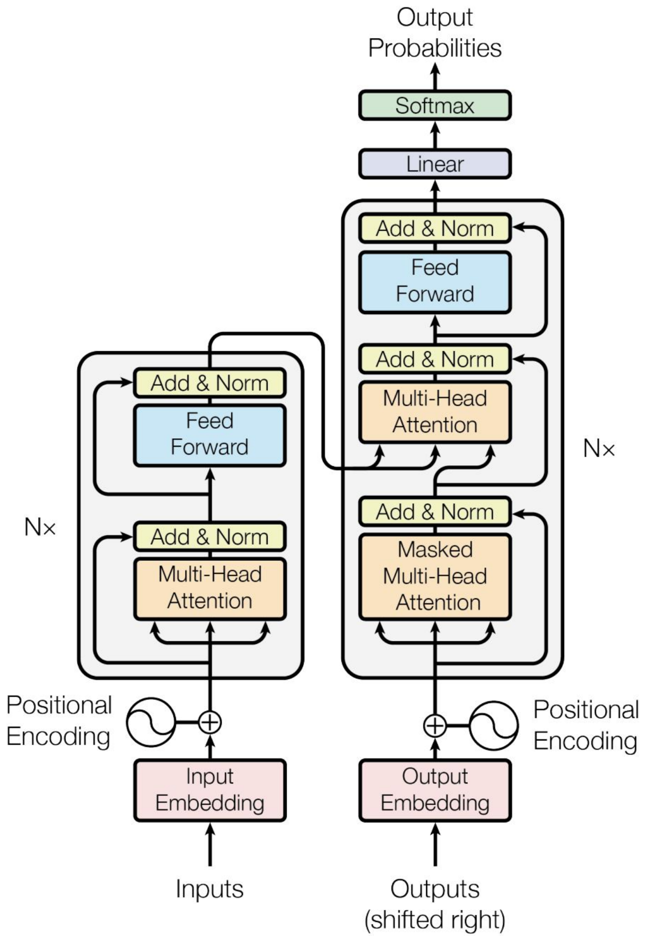


Figure 1: The Transformer - model architecture.

- **RoPE嵌入（Rotary Position Embedding）**

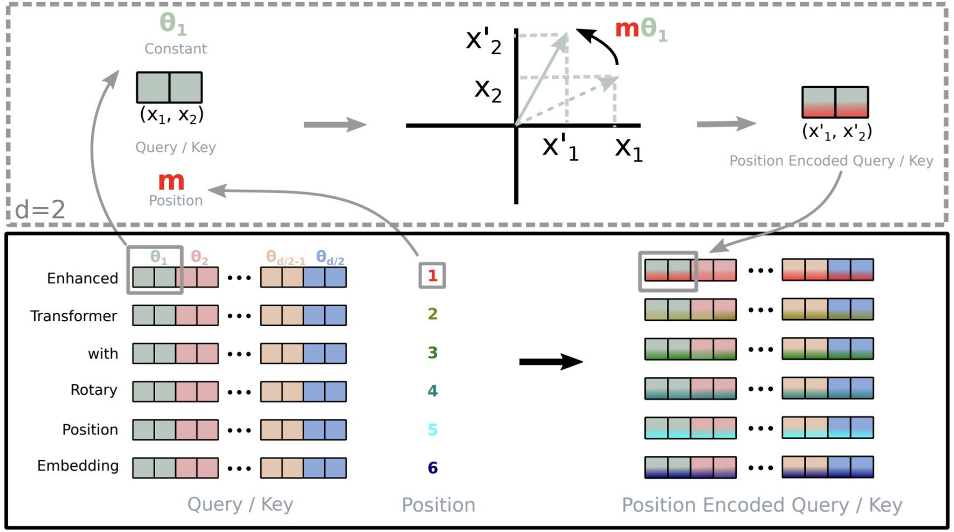


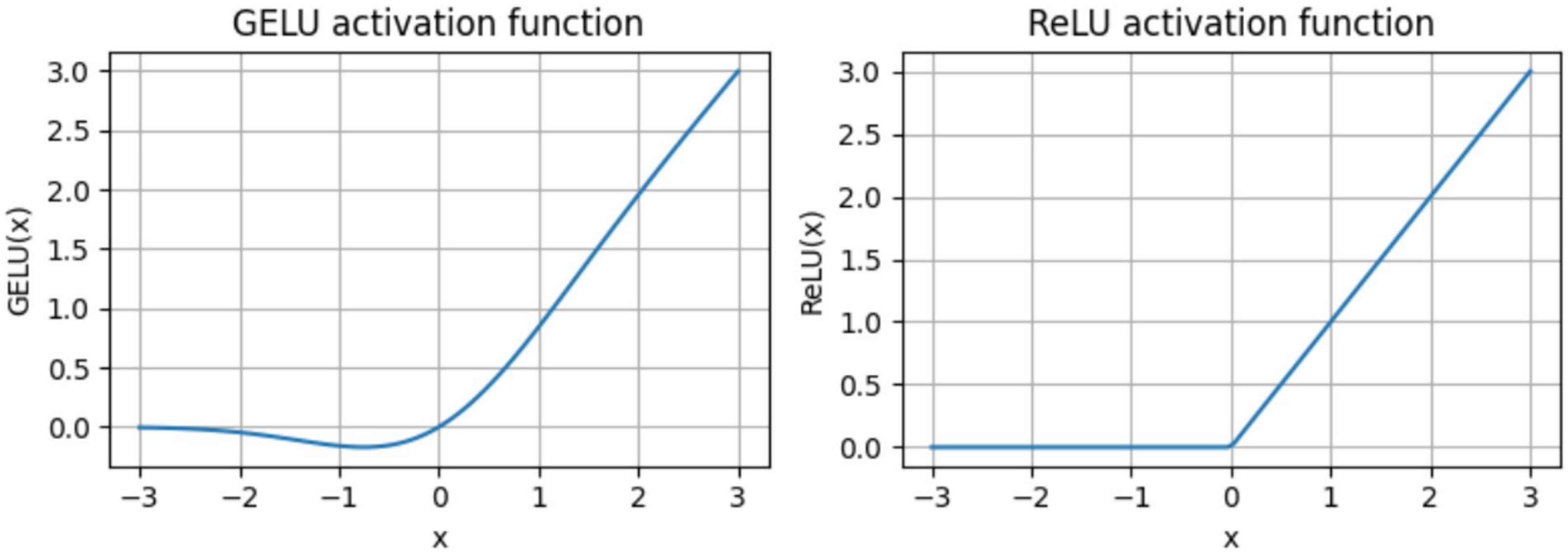
Figure 1: Implementation of Rotary Position Embedding(RoPE).

左图来源：Attention Is All You Need

右图来源：ROFORMER: ENHANCED TRANSFORMER WITH ROTARY POSITION EMBEDDING

模型结构

- **GeGLU激活函数**



图片来源：Understanding, Using, and Finetuning Gemma

模型结构

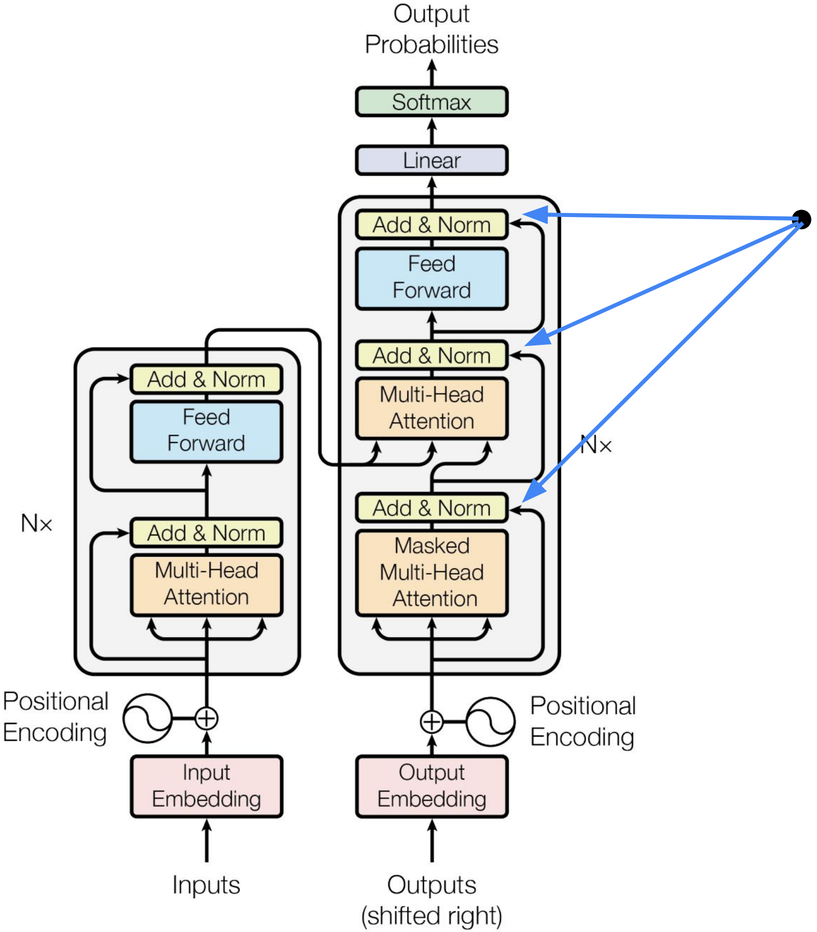
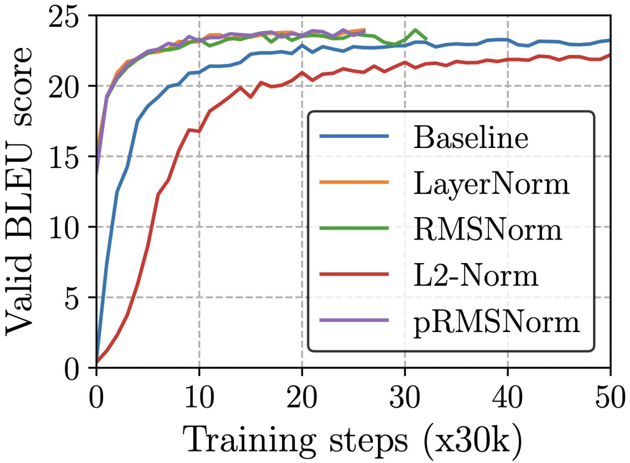


Figure 1: The Transformer - model architecture.

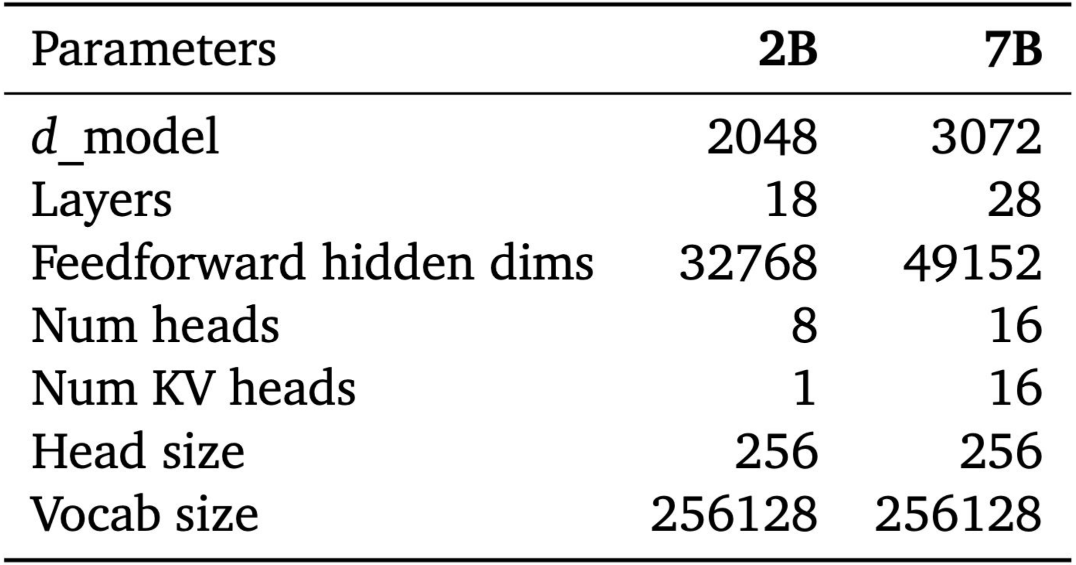
RMSNorm

$$\mathrm{RMSNorm}(x_i) = \frac{x_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 + \epsilon}}$$



左图来源: Attention Is All You Need 右图来源: Root Mean Square Layer Normalization

模型结构



Parameters	**2B**	**7B**
*d*_model	2048	3072
Layers	18	28
Feedforward hidden dims	32768	49152
Num heads	8	16
Num KV heads	1	16
Head size	256	256
Vocab size	256128	256128

Table 1 | Key model parameters.

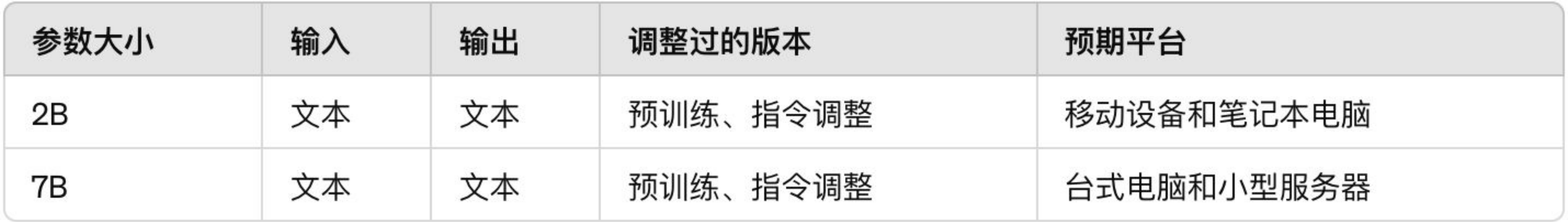
表格来源：[Gemma: Open Models Based on Gemini Research and Technology]()

训练数据

- Gemma 2B和7B分别在3万亿和6万亿的主要以英语为主的数据上进行了训练
- 数据来自网络文档、数学和编程内容
- Gemma模型不是多模态的，也没有针对多语言任务的最新表现进行训练
- 词汇表的大小为256,000个token

信息来源：Gemma: Open Models Based on Gemini Research and Technology

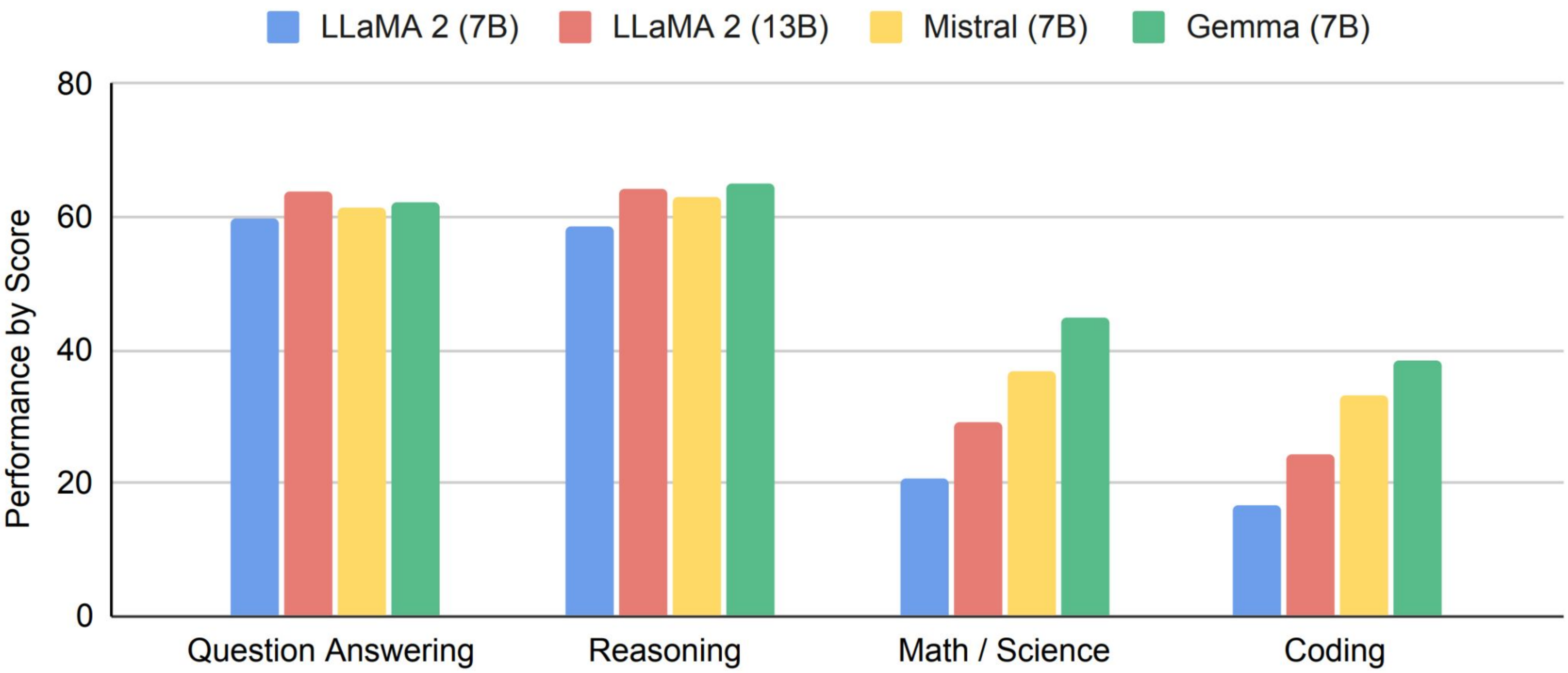
模型尺寸与性能



参数大小	输入	输出	调整过的版本	预期平台
2B	文本	文本	预训练、指令调整	移动设备和笔记本电脑
7B	文本	文本	预训练、指令调整	台式电脑和小型服务器

表格来源：Gemma: Open Models Based on Gemini Research and Technology

模型性能



图片来源：Gemma: Open Models Based on Gemini Research and Technology

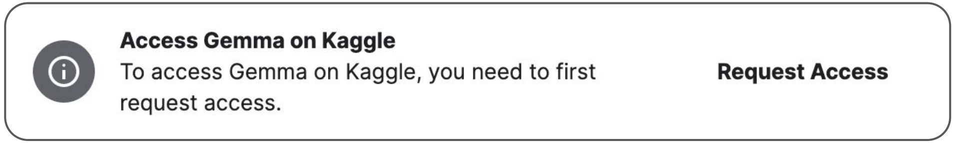
Build
with AI

Google Developer Groups
天津

开始使用Gemma

获取Gemma模型

kaggle



Access Gemma on Kaggle
To access Gemma on Kaggle, you need to first request access.

Request Access

准备工作

首先，将Keras 3和KerasNLP安装到您的环境中，然后导入keras_nlp模块。

```
!pip install --upgrade keras-nlp
!pip install --upgrade keras

import keras_nlp
```

接着，从预设配置中加载Gemma模型！

```
# https://keras.io/api/keras_nlp/models/gemma/gemma_causal_lm/
g_lm = keras_nlp.models.GemmaCausalLM.from_preset("gemma_2b_en")
```

预设配置可用于Gemma的2B和7B参数版本。

使用Gemma

只需将提示词传递给**generate()**函数，并可选地指定响应的最大长度。

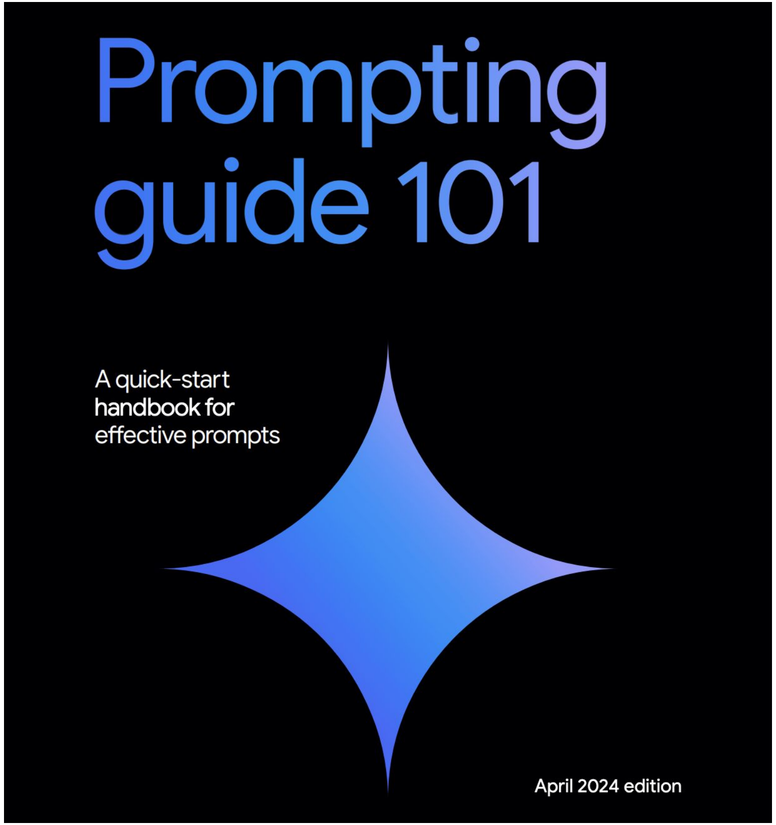
例如，如果问Gemma“it was a dark and stormy night.”

```
txt = g_lm.generate("It was a dark and stormy night.", max_length=64)
print(txt)
```

It was a dark and stormy night.
The rain was pouring down, and the wind was howling.
But that didn’t stop a group of friends from going out for a night of fun.
They were all dressed up in their best clothes, and they were ready to have a good time.

提示词编写指导

Prompting Guide 101



Prompting guide 101

A quick-start handbook for effective prompts

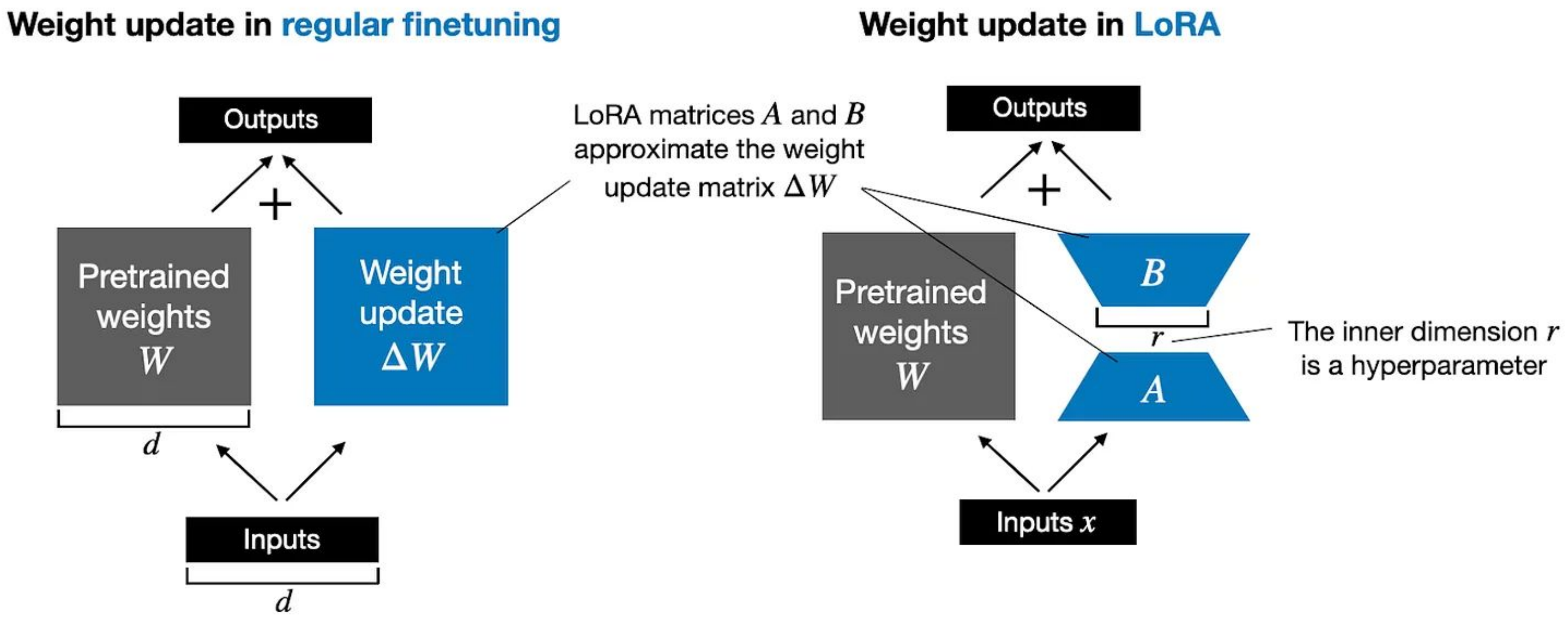
April 2024 edition

微调Gemma

- 使用Keras进行微调
- 支持使用LoRA进行微调

K Keras

LoRA


Weight update in regular finetuning
Outputs
+
Pretrained weights W
Weight update ΔW
d
Inputs
d
LoRA matrices A and B approximate the weight update matrix ΔW
Weight update in LoRA
Outputs
+
Pretrained weights W
B
r
A
The inner dimension r is a hyperparameter
Inputs x

Google Developer Groups

图片来源 : Practical Tips for Finetuning LLMs Using LoRA (Low-Rank Adaptation)

微调Gemma

- 使用Keras进行微调
- 支持使用LoRA进行微调

Keras

```
gemma.backbone.enable_lora(rank=8)
# fine-tune ...
gemma.fit(...)
gemma.backbone.save_lora_weights("lora.h5")
```

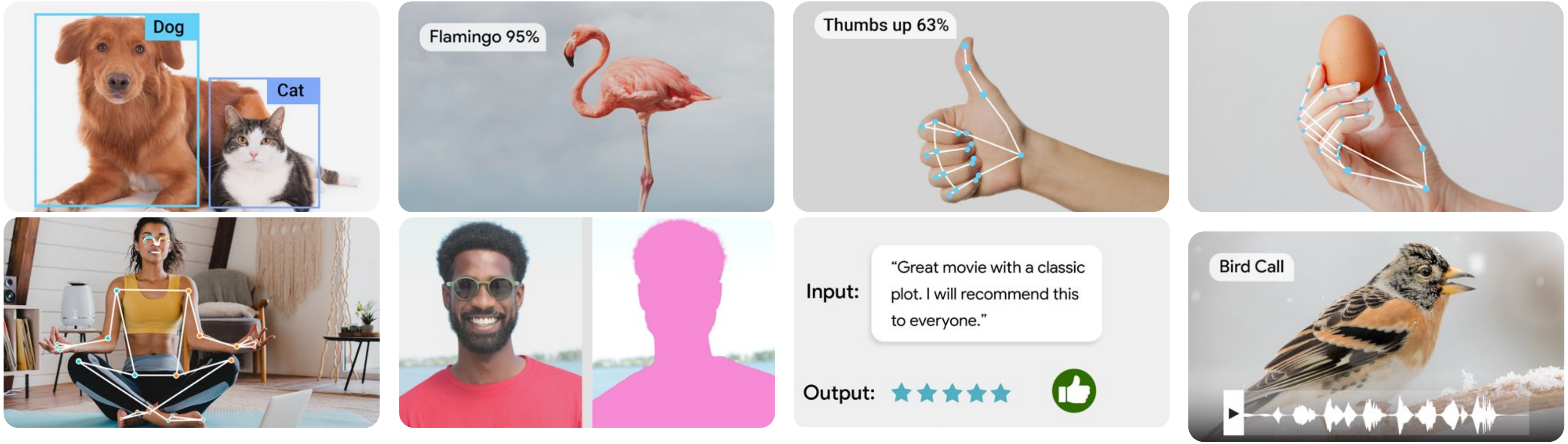
分布式微调Gemma

```
devices=keras.distribution.list_devices()
device_mesh = keras.distribution.DeviceMesh((1, 8),["batch", "model"], devices))
layout_map = keras.distribution.LayoutMap(device_mesh)

# Partitioning for embeddings (regex)
layout_map["token_embedding/embeddings"] = (None, "model")
# Partitioning (regex) for attention layer weights
layout_map["decoder_block.*attention.*(query|key|value).*kernel"] = (None, "model", None)
layout_map["decoder_block.*attention_output.*kernel"] = (None, None, "model")
layout_map["decoder_block.*ffw_gating.*kernel"] = ("model", None)
layout_map["decoder_block.*ffw_linear.*kernel"] = (None, "model")

keras.distribution.set_distribution(keras.distribution.ModelParallel(device_mesh,
                                                                     layout_map,
                                                                     batch_dim_name="batch"))

# - load the model here
```

微调Gemma

- 使用Keras进行微调
- 支持使用LoRA进行微调
- 支持JAX进行微调

使用MediaPipe在设备上的集成Gemma


Dog
Cat
Flamingo 95%
Thumbs up 63%
Input:
"Great movie with a classic plot. I will recommend this to everyone."
Output: ★★★★★
Bird Call

- Studio: https://mediapipe-studio.webapps.google.com/demo/llm_inference
- Documentation: https://developers.google.com/mediapipe/solutions/genai/llm_inference

代码示例

- Android:
  - https://github.com/googlesamples/mediapipe/tree/main/examples/llm_inference/android
  - https://github.com/NSTiwari/Gemma-on-Android
- Web: https://github.com/googlesamples/mediapipe/tree/main/examples/llm_inference/js
- iOS: https://github.com/googlesamples/mediapipe/tree/main/examples/llm_inference/ios

Build with AI

Google Developer Groups
天津

Gemma新进展

Gemma家族

- CodeGemma
- PaliGemma（中文介绍）
- RecurrentGemma

参考资料

1.Gemma的官方网站：https://ai.google.dev/gemma/docs

2.Gemma技术报告
：https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf

3.Gemma: Introducing new state-of-the-art open model by Google, https://medium.com/@shravankoninti/gemma-introducing-new-state-of-the-art-open-model-by-google-caae9fe29972

4.Understanding, Using, and Finetuning Gemma, https://lightning.ai/lightning-ai/studios/understanding-using-and-finetuning-gemma

5.What is Low-Rank Adaptation (LoRA) | explained by the inventor, https://www.youtube.com/watch?v=DhRoTONcyZE

Build with AI

码出未来，现在开始！