专家混搭

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed

摘要

我们介绍了Mixtral 8x7B，这是一种稀疏专家混合模型（SMoE）的语言模型。Mixtral与Mistral 7B具有相同的架构，不同之处在于每个层由8个前馈块组成（即专家）。对于每一个令牌，在每一层中，一个路由器网络选择两个专家来处理当前状态，并将它们的输出进行组合。尽管每个令牌只看到两个专家，但在每一步的时间步中，所选的专家可能会有所不同。因此，每个令牌可以访问47亿个参数，但仅在推理过程中使用13亿个活跃参数。Mixtral通过训练集大小为32k令牌的方式进行了训练，并且在所有评估基准上超过了或匹配了Llama 2 70B和GPT-3.5。特别是，在数学、代码生成和多语言基准中，Mixtral远超Llama 2 70B。我们还提供了一个模型微调版本，名为Mixtral 8x7B – Instruct，该模型在遵循指令的人类基准上超越了GPT-3.5 Turbo、Claude-2.1和Gemini Pro以及Llama 2 70B – chat模型。两个基础模型和微调模型都以Apache 2.0许可发布。

Code: https://github.com/mistralai/mistral-src Webpage: https://mistral.ai/news/mixtral-of-experts/

1 引言

在本论文中，我们提出了一种稀疏专家混合模型（SMoE），该模型具有开放权重，并且受 Apache 2.0 许可证的约束。Mixtral 在大多数基准测试上优于 Llama 2 70B 和 GPT-3.5。由于它仅使用每 token 的一小部分参数，Mixtral 可以在低批次大小时提供更快的推理速度，并且在大规模批次大小时具有更高的吞吐量。

Mixtral 是一种稀疏混合专家网络。它是解码器模型，其中前向块从一组 8 个不同的参数组中选择。在每一层中，对于每个令牌，路由器网络会选择两个这些组（称为“专家”）来处理该令牌并将其输出相加。这种技术增加了模型的参数数量，同时控制成本和延迟，因为模型仅使用每令牌总参数集的少量部分。

Mixtral 是通过使用大小为 32K 块的上下文数据进行预训练的。它要么匹配，要么超过 LLAMA 2 70B 和 GPT-3.5 在多个基准测试中的表现。特别是，

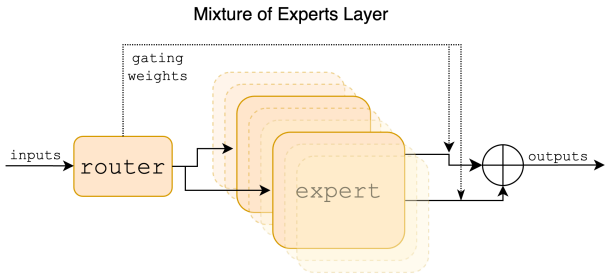

图1：专家混合层。每个输入向量由路由器分配到8个专家中的2个。该层的输出是选择的两个专家输出的加权和。在Mixtral中，专家是一个类似于标准前馈块的自回归架构中的普通变压器架构。

Mixtral 在数学、代码生成以及需要多语言理解的任务中展示了卓越的能力，显著优于 Llama 2 70B 在这些领域。实验表明，Mixtral 能够成功从其 32k 个令牌的上下文中检索信息，无论序列长度如何，信息在序列中的位置如何。

我们还展示了 Mixtral 8x7B – Instruct，这是一种通过监督微调和直接偏好优化（[25]）进行微调的聊天模型。其性能显著超过了 GPT-3.5 Turbo、Claude-2.1、Gemini Pro 和 Llama 2 70B——在人类评估基准上的聊天模型。Mixtral – Instruct 还展示了减少的偏见，并且在基准测试中，如 BBQ 和 BOLD 中具有更平衡的情感倾向。

我们同时发布了Mixtral 8x7B和Mixtral 8x7B – Instruct，均在Apache 2.0许可下[1]下提供，免费供学术和商业用途使用，确保了广泛的可访问性和多样化的应用潜力。为了使社区能够以完全开源的堆栈运行Mixtral，我们向vLLM项目提交了更改，该项目整合了Megablocks CUDA核函数以实现高效推理。Skypilot还允许在云上的任何实例上部署vLLM端点。

2 建筑细节

Mixtral 是基于变压器架构 [31] 开发的，并且使用了与文献 [18] 中描述相同的变化，但值得注意的是 Mixtral 支持一个完全密集的上下文长度为 32k 的令牌数，并且前向层块被混合专家层（第 2.1 节）所取代。模型架构参数总结见表 1。

Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
context_len	32768
vocab_size	32000
num_experts	8
top_k_experts	2

表 1：模型架构。

2.1 稀疏专家混合体

我们对混合专家层（图1）进行了简要概述。对于更深入的概述，请参见[12]。给定输入 x，MoE 模块的输出由专家网络的输出按权重之和确定，其中权重由门控网络的输出给出。即，给定 n 专家网络 $\{E_0, E_i, ..., E_{n-1}\}$，专家层的输出为：

$$\sum_{i=0}^{n-1} G(x)_i \cdot E_i(x).$$

这里，$G(x)_i$ 表示门控网络对第 i- 专家的 n- 维输出，而 $E_i(x)$ 是第 i- 专家网络的输出。如果门控向量稀疏，则我们可以避免计算那些门控为零的专家的输出。$G(x)$ 的实现方式有多种选择 [6, 15, 35]，但简单且高效的实现是通过对线性层的 Top-K 特征值进行 Softmax 实现的 [28]。我们使用

$$G(x) := \text{Softmax}(\text{TopK}(x \cdot W_g)),$$

where $(\text{TopK}(\ell))_i$ if ℓ_i is among the top-K coordinates of logits $\ell \in \mathbb{R}^n$ and $(\text{TopK}(\ell))_i$ otherwise. The value of K – the number of experts used per token – is a hyper-parameter that modulates the amount of compute used to process each token. If one increases n while keeping K fixed, one

[1] https://mistral.ai/news/mixtral-of-experts/

可以增加模型的参数数量，同时保持其计算成本基本恒定。这促使我们区分了模型的总参数数量（通常称为稀疏参数数量），随着 n 增长；以及处理单个词元所需的参数数量（被称为活跃参数数量），随着 K 到 n 的增长。

MoE 层可以在单个 GPU 上高效运行，性能优异的专门化核。例如，Megablocks [13] 将 MoE 层中的前向传播网络（FFN）操作转换为大规模稀疏矩阵乘法，显著提高了执行速度，并自然地处理了不同专家分配给它们的令牌数量变量的情况。此外，MoE 层可以通过标准模型并行技术分布在多个 GPU 上，并通过特定类型的分区策略称为专家并行（EP） [28]。在 MoE 层执行过程中，旨在由特定专家处理的令牌被路由到相应的 GPU 进行处理，并且专家的输出返回到原始令牌位置。需要注意的是，EP 引入了负载平衡挑战，因为它必须确保均匀分配工作量以防止单个 GPU 超载或遇到计算瓶颈。

在 Transformer 模型中，MoE 层独立应用于每个令牌，并替换变压器块中的前馈 (FFN) 子模块。对于 Mixtral，我们使用与专家函数相同的 SwiGLU 架构 $E_i(x)$ 并设置 $K = $ 为 2。这意味着每个令牌被路由到两个具有不同权重集的 SwiGLU 子模块中。将所有这些因素结合起来，输入令牌 x 的输出 y 计算如下：

$$y = \sum_{i=0}^{n-1} \text{Softmax}(\text{Top2}(x \cdot W_g))_i \cdot \text{SwiGLU}_i(x).$$

该公式与 GShard 架构类似 [21]，区别在于我们用 MoE 层替换了所有 FFN 子块，而 GShard 则将每个其他块替换为其他块，并且 GShard 对第二专家分配给每个令牌的门控策略更为复杂。

三结果

我们对比Mixtral与Llama，并使用我们的评估管道重新运行所有基准测试，以实现公平的比较。我们将性能测量在以下分类的任务上进行：

- 常识推理 (零样本): Hellaswag [32]，Winogrande [26]，PIQA [3]，SIQA [27]，OpenbookQA [22]，ARC-Easy，ARC-Challenge [8]，CommonsenseQA [30]
- 世界知识（5次样本）：自然问题 [20]， triviaqa [19]
- 阅读理解（零样本学习）：BoolQ [7]，QuAC [5]
- 数学：GSM8K [9]（8发）与maj@8和MATH [17]（4发）与maj@4
- 代码：Humaneval [4]（零样本）和MBPP [1]（三样本）
- 热门聚合结果：MMLU [16]（5-样本），BBH [29]（3-样本），以及AGI 评估 [34]（3-5-样本，仅限英语多项选择题）

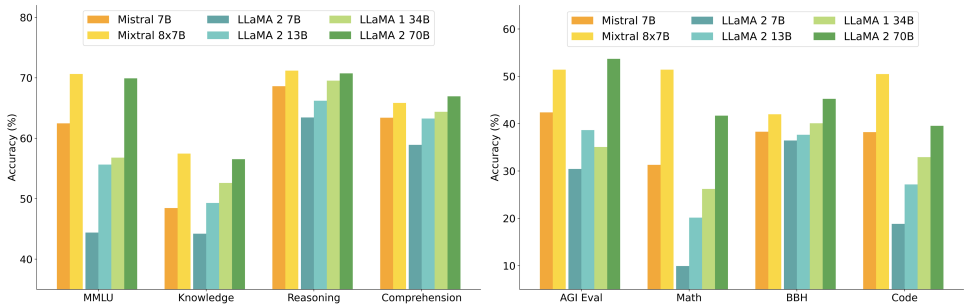

图 2：Mixtral 在广泛基准上的性能表现。所有模型均在我们的评估管道上重新评估了所有指标，以实现准确的比较。 Mixtral 在所有基准测试中都优于或与 Llama 2 70B 相媲美。特别是，在数学和代码生成方面，它具有显著优势。

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c	NQ	TriQA	HumanE	MBPP	Math	GSM8K
LLaMA 2 7B	7B	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	17.5%	56.6%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	13B	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	16.7%	64.0%	18.9%	35.4%	6.0%	34.3%
LLaMA 1 33B	33B	56.8%	83.7%	76.2%	82.2%	79.6%	54.4%	24.1%	68.5%	25.0%	40.9%	8.4%	44.1%
LLaMA 2 70B	70B	69.9%	**85.4%**	**80.4%**	82.6%	79.9%	56.5%	25.4%	**73.0%**	29.3%	49.8%	13.8%	69.6%
Mistral 7B	7B	62.5%	81.0%	74.2%	82.2%	80.5%	54.9%	23.2%	62.5%	26.2%	50.2%	12.7%	50.0%
Mixtral 8x7B	13B	**70.6%**	84.4%	77.2%	**83.6%**	**83.1%**	**59.7%**	**30.6%**	71.5%	**40.2%**	**60.7%**	**28.4%**	**74.4%**

表 2：Mixtral 与 Llama 的比较。在几乎所有流行的基准测试中，Mixtral 能够超过或等于 Llama 2 70B 的性能，同时在推理过程中使用了 5 倍于参数的少。

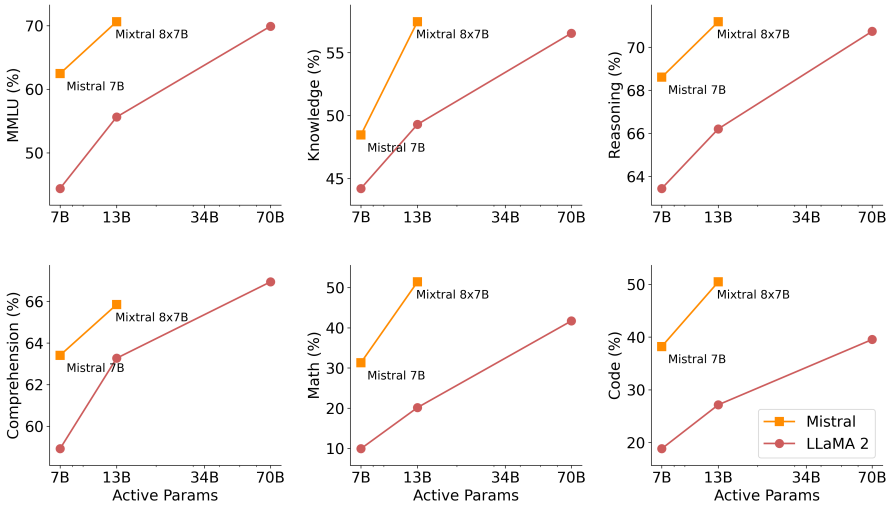

Figure 3：Mistral（7B/8x7B）与Llama 2（7B/13B/70B）在MMLU、常识推理、世界知识和阅读理解、数学和代码方面的结果。Mistral在所有基准测试中均优于Llama 2 70B，除了阅读理解基准外，使用了5倍更低的活跃参数。此外，在代码和数学方面也远胜于Llama 2 70B。

详细结果报告了 Mixtral、Mistral-7B 和 LLaMA-2 7B/13B/70B/LaMa1 34B[2] 的详细结果。在表 2 中，图 2 比较了 Mixtral 在不同类别中的性能与 LLaMA 模型的对比情况。Mixtral 在大多数指标上超越了 LLaMA-2 70B。特别是，在代码和数学基准测试中，Mixtral 显示出更优异的表现。

大小和效率。我们与Llama 2家族进行比较，旨在理解Mixtral模型在成本-性能谱中的效率（参见图3）。作为稀疏专家混合模型，Mixtral每个令牌仅使用13亿个活跃参数。通过减少5倍的活跃参数，Mixtral能够在大多数类别中超越Llama 2 70B。

注意，本分析关注的是活跃参数计数（见第2.1节），与推理计算成本成正比，但不考虑内存成本和硬件利用率。服务Mixtral的内存成本与其*sparse*参数数量相关，即47B，仍小于Llama 2 70B。至于设备利用率，我们注意到SMoEs层由于路由机制引入额外开销，并且当在一个设备上运行多个专家时，内存负载增加导致硬件利用率更高。它们更适合批处理工作负载，在这种情况下可以达到良好的算术强度。

比较与 LLaMA 2 70B 和 GPT-3.5。在表 3 中，我们报告了 Mixtral 8x7B 在与 LLaMA 2 70B 和 GPT-3.5 的性能对比中的表现。观察到 Mixtral 的表现与其其他两个模型相当或更好。在 MMLU 上，Mixtral 获得了更好的性能，尽管其容量显著较小（47B 令牌对 70B）。对于 MT Bench，我们报告了最新可用的 GPT-3.5-Turbo 模型的表现，gpt-3.5-turbo-1106。

[2]Since Llama 2 34B was not open-sourced, we report results for Llama 1 34B.