**Mixtral of Experts**

专家混合模型

**太长不看版**

本文介绍了一种稀疏专家混合语言模型Mixtral 8x7B。该模型每个层由八个专家构成，每个token通过路由被发送到两个专家并组合输出，这样每个token推断期间只激活部分参数。模型在32k tokens的前后文下训练后与Llama 2 70B及GPT-3.5性能持平，甚至在数学，代码生成和多语言基准测试中占极大优势。作者同时还提供了一个经过指令微调的模型Mixtral 8x7B – Instruct，超越了当前大量主流LLM。

**摘要**

本文介绍了一种稀疏专家混合（Sparse Mixture of Experts， SMoE）语言模型Mixtral 8x7B。 该模型每个层由八个前馈块（专家）构成，每个token通过路由被发送到两个专家并组合输出。因所选专家在不同时间步可能不同，虽然每个token可以访问47B的参数，推断期间只激活13B的参数。作者在32k tokens的前后文下对模型进行了训练，结果与Llama 2 70B及GPT-3.5匹敌，并在数学，代码生成和多语言基准测试中远远优于两者。作者同时还提供了一个经过微调以遵循指令的模型Mixtral 8x7B – Instruct，超越了GPT-3.5 Turbo、Claude-2.1、Gemini Pro和Llama 2 70B – chat。

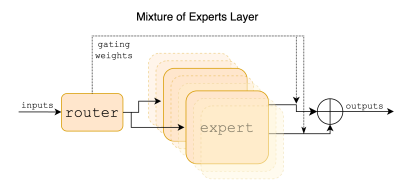


图 1: 专家混合层。 每个输入向量通过路由器分配给8个专家中的2个。该层的输出是所选两个专家输出的加权和。在Mixtral中，专家是一个标准的前馈块，就像vanilla transformer架构中的一样。

**1引言**

Mixtral 8x7B是一个稀疏的专家混合模型。在大多数基准测试中，Mixtral的表现优于 Llama 2 70B 和 GPT- 3.5。由于它仅针对每个token使用其参数的子集，Mixtral可以在小批量大小下实现更快的推理速度，并在大批量大小下实现更高的吞吐量。

Mixtral是一个解码器模型，前馈块从8 个不同的参数组中选择。在每一层中，路由器网络选择其中2组（即“专家”）来处理每个token并将它们的输出进行加法组合。这种技术增加了模型的参数数量，同时控制成本和延迟。

Mixtral使用多语言数据进行预训练。在多个基准测试中，它与 Llama 2 70B 和 GPT-3.5 的性能相匹敌甚至超越它们，特别是在数学、代码生成以及需要多语言理解的任务中。实验证明，Mixtral能够成功地从其 32k 个token的上下文窗口中检索信息，无论序列长度和信息在序列中的位置如何。

作者还介绍了Mixtral 8x7B – Instruct，这是一个经过监督微调和直接偏好优化的指令跟随聊天模型。其性能显著超越了 GPT-3.5 Turbo、Claude-2.1、Gemini Pro 和 Llama 2 70B –chat在人类评估基准测试中的表现，展示出更少的偏见和更平衡的情感特征。

作者在 Apache 2.0 许可下发布了 Mixtral 8x7B和 Mixtral 8x7B – Instruct。为了使社区能够使用完全开源堆栈运行Mixtral，作者向集成了 Megablocks CUDA 内核的vLLM 项目提交了更改。Skypilot 还允许在云中的任何实例上部署 vLLM 端点。

**2 结构细节**

该模型基于transformer架构，采用与mistral 7b相同的修改，但前馈层被专家混合层取代，且支持完全密集的32k tokens的上下文长度。模型架构参数总结在表1中。

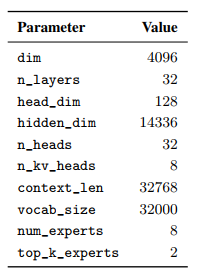


表1: 模型结构。

**2.1 专家的稀疏混合**

对于输入x，MoE模块的输出由专家网络输出的加权和确定，其中权重由门控网络的输出给出。即给定*n*个专家网络，专家层的输出为：

这里，表示第 *i* 个专家的门控网络的 *n* 维输出，是第 *i* 个专家网络的输出。由于门控向量是稀疏的，可以免去计算门控为零的专家的输出。实现 的简单高效的方法是对线性层的前 K 个 logit 进行 softmax 处理。即

*G*(*x*) := Softmax(TopK(*x · Wg*))

这里，(TopK(*ℓ*))*i* := *ℓi* 表示如果 *ℓi* 在 logits *ℓ ∈* 的前 K 个坐标中，则 (TopK(*ℓ*))*i* := *ℓi*，否则 (TopK(*ℓ*))*i* := *-∞*。每个token使用的专家数量K是一个超参数。保持 K 不变增加 *n*，可以增加模型的参数数量，同时保持其计算成本基本恒定。这里要区分模型的总参数数量（通常称为稀疏参数数量），它随着 *n* 的增长而增长，以及用于处理单个token的参数数量（称为活跃参数数量），它随着K增长到*n 。*

MoE 层可以在单个具有高性能专用内核的GPU 上高效运行。例如，Megablocks 将MoE 层的FFN操作转换为大型稀疏矩阵乘法，显著提高执行速度，并能处理不同专家获得的token数量不同的情况。此外，通过一种称为专家并行（EP）的特定分区策略，MoE 层被分布到多个 GPU 上，这样分配给特定专家的token被路由到相应的 GPU 进行处理，并将专家的输出返回到原始token位置。但EP在负载平衡方面存在挑战。

在 Transformer 模型中，MoE 层独立应用于每个token，并替换了前馈（FFN）子块。对于 Mixtral，作者使用与专家函数 *Ei*(*x*) 相同的 SwiGLU 架构，并设置 *K* = 2，使每个token被路由到两个具有不同权重集的 SwiGLU 子块。对于输入token *x*，输出 *y* 如下：

该公式与GShard架构类似，但GShard替换了其他块而非FFN，并为每个token分配的第二个专家使用了更精细的门控策略。

**3结果**

作者使用自己的评估流程重新运行所有基准测试，以将Mixtral与Llama进行公平比较。各种任务的性能分类如下：

• 常识推理 （**0-shot**） **:** Hellaswag，Winogrande，PIQA，SIQA，OpenbookQA，ARC-Easy，ARC-Challenge，CommonsenseQA

• 世界知识（**5-shot**）**:** NaturalQuestions，TriviaQA

• 阅读理解（**0-shot**）**:** BoolQ，QuAC

• 数学**:**使用 maj@8的GSM8K（8-shot），使用 maj@4的MATH（4-shot）

• 代码**:** Humaneval（**0-shot**），MBPP（3- shot）

• 热门的聚合结果**:** MMLU（5-shot），BBH（3- shot），AGI Eval（3-5-shot，仅英文多项选择题）

模型比较的详细结果见表 2。

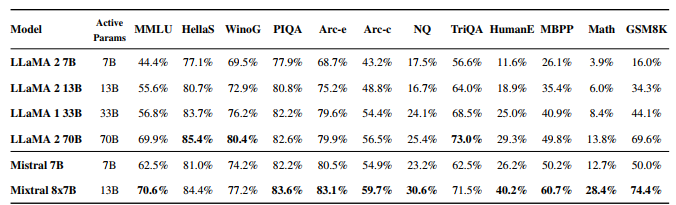


表2: 与Llama的比较。

图 2 比较了 Mixtral在不同类别中与 Llama模型的性能。 Mixtral在大多数指标上超过了 Llama 2 70B，特别是在代码和数学基准测试中。

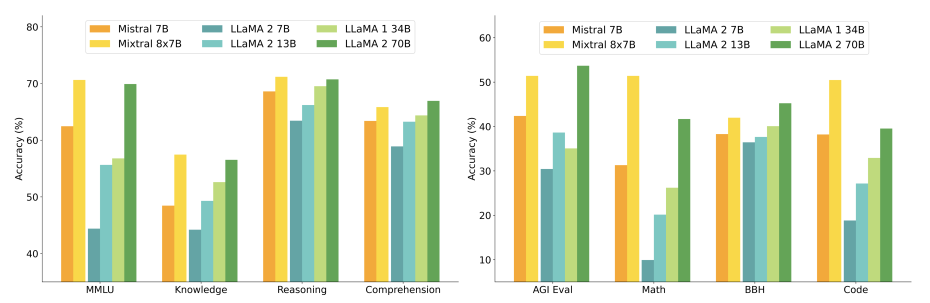


图2: 在广泛基准测试中Mixtral和不同的Llama模型的性能。

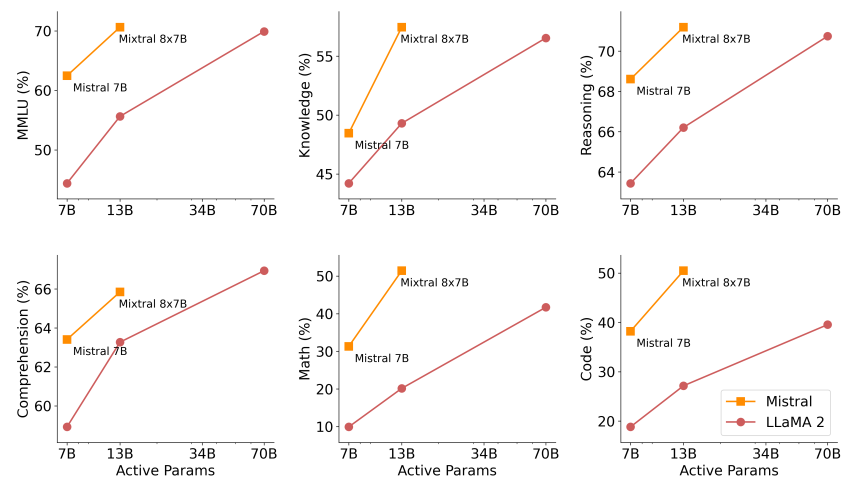


图 3: MMLU、常识推理、世界知识和阅读理解、数学和代码上，Mistral（7B/8x7B）vs Llama 2（7B/13B/70B）的结果。

规模和效率。为了解 Mixtral模型在成本-性能谱中的效率，作者将模型与 Llama 2 系列进行比较（见图 3）。不考虑内存成本和硬件利用率的情况下，活跃参数数量与推理计算成本成正比。Mixtral每个token仅使用比 Llama 2 70B 低 5 倍的 13B 个活跃参数，能够在大多数类别上胜过 Llama 2 70B。为 Mixtral提供服务的内存成本与其稀疏参数47B成正比，仍小于 Llama 2 70B。设备利用率方面，由于路由机制和增加的内存负载，当每个设备运行多个专家时，SMoEs层引入了额外的开销。它们更适合批处理工作负载，以达到良好的计算强度。

与 **Llama 2 70B** 和 **GPT-3.5** 的比较。作者从表 3观察到 Mixtral的性能与另外两个模型相似或更好。

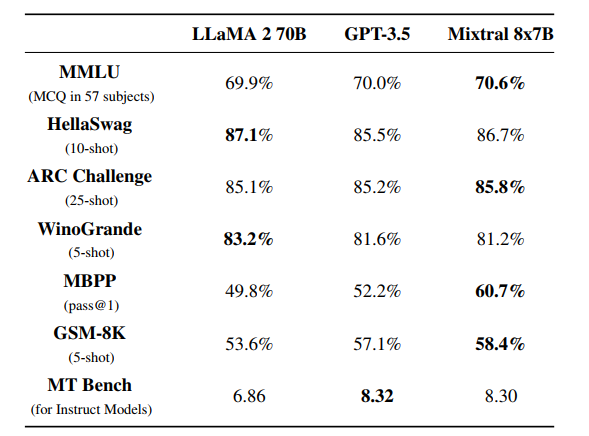


表 3: 与 Llama 2 70B 和 GPT-3.5 的比较。

评估差异。 在一些基准测试中，作者的评估协议与 Llama 2 论文中报告的协议存在差异：1）在 MBPP 上使用手动验证的子集；2）在 TriviaQA 上不提供维基百科上下文。

**3.1 多语言基准测试**

与 Mistral 7B相比，Mixtral在预训练过程中显著增加了多语言数据的比例，使其在多语言基准测试中表现良好，同时在英语中保持高准确性。如表 4所示，Mixtral在法语、德语、西班牙语和意大利语方面显著优于Llama 2 70B。

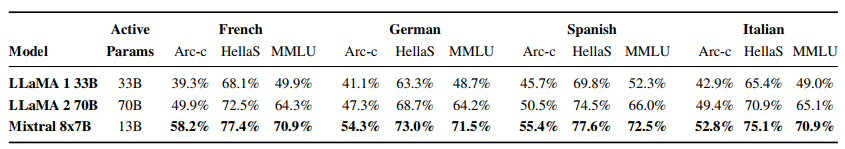


表 4: 多语言基准测试上Mixtral与 Llama的比较。

**3.2长范围表现**

作者在密码检索任务上对其进行评估，这是一个用来衡量模型在长提示中随机插入密码时检索能力的合成任务。图4（左）中的结果显示，无论上下文长度或密码在序列中的位置如何，Mixtral都实现了100% 的检索准确率。图 4（右）显示，Mixtral在证据堆数据集子集上的困惑度(perplexity)随着上下文大小增加而递减。

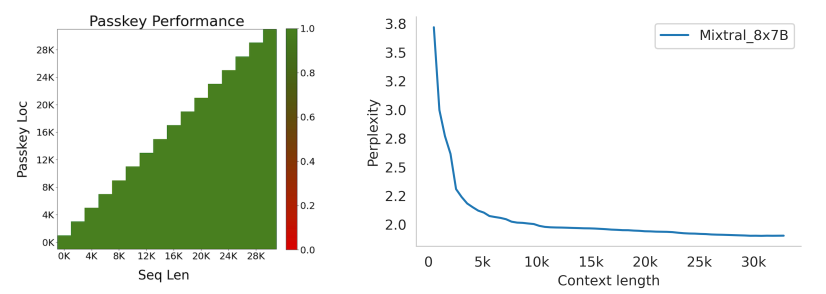


图 4: Mixtral的长距离性能。

**3.3 偏见基准测试**

为了识别可能需要通过微调/偏好建模进行修正的缺陷，作者在问答偏见基准（BBQ）和开放式语言生成数据集中的偏见（BOLD）上测量基础模型的性能。BBQ 是一个针对九个不同的社会偏见类别的手写问题数据集，BOLD 是一个包含五个领域的英文文本生成提示的大规模数据集。

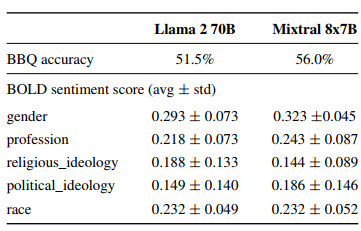


图 5: 偏见基准测试。

作者自用评估框架在 BBQ 和BOLD 上对 Llama 2 和 Mixtral进行基准测试，结果见表 5。Mixtral在 BBQ 基准测试上表现出更少的偏见（56.0% vs 51.5%）。BOLD 中，Mixtral展示出更高的平均情感分数，意味着更积极的情感，以及较低的标准偏差，表示组内的偏见较少。

**4 指令调优**

作者使用监督微调（SFT）在一个指导数据集上训练 Mixtral – Instruct，然后在一个配对反馈数据集上进行直接偏好优化（DPO）。 模型在 MT-Bench 上达到 8.30 的分数（见表 2），成为截至 2023 年 12 月最佳的开放权重模型。由 LMSys 进行的独立人类评估报告如图 6。

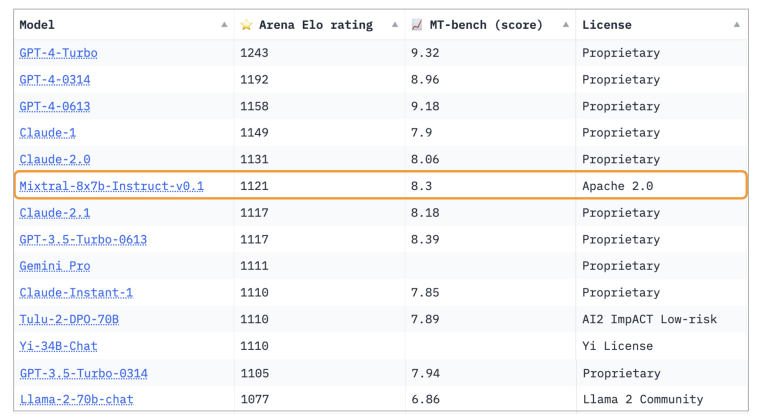


图 6: LMSys 排行榜. Mixtral 8x7B指导 v0.1 实现了1121的竞技场 Elo 评分，超过了 Claude-2.1（1117）、所有版本的 GPT-3.5-Turbo（1117 最佳）、Gemini Pro（1111）和 Llama-2-70bchat（1077）。

**5路由分析**

本节对路由器进行了有关专家选择的小型分析。作者尤其感兴趣的是训练中是否有些专家专门针对某些特定领域。

作者测量了在The Pile验证数据集的不同子集上选择的专家的分布。对于0、15和31层（分别是模型的第一层和最后一层）的结果见图 7。令人惊讶的是，并未观察到明显的基于主题的专家分配模式。在所有层中，对于ArXiv论文（使用Latex编写）、生物学（PubMed摘要）和哲学（PhilPapers）文档的专家分配非常相似。

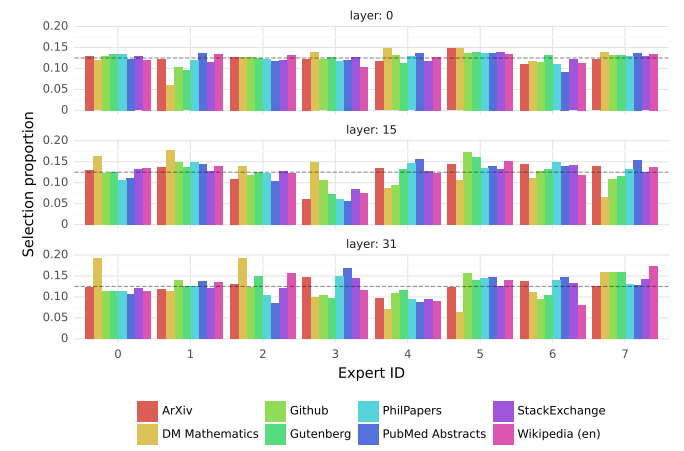


图 7: The Pile数据集中不同领域中每个专家被分配的token比例。灰色虚线垂直线token为1/8，即均匀抽样时的预期比例。这里考虑作为路由器的第一或第二选择的专家。

仅对于DM数学，专家的分布略有不同。这种分歧可能是数据集合成性质和对自然语言范围的覆盖有限的结果，尤其在第一层和最后一层中，隐藏状态与输入输出非常相关。

这表明路由器确实表现出一些结构化的句法行为。图8展示了来自不同领域的文本示例。在Python中的‘self’和英语中的‘Question’这样的词往往会路由到相同的专家，即使涉及多个token。同样，代码中缩进token总是分配给相同的专家，尤其在第一层和最后一层，隐藏状态更与模型的输入和输出相关。



图 8: 每个token都用第一个专家选择的颜色着色的文本样本。

作者还注意到，连续的token通常被分配给相同的专家。The Pile数据集中存在一定程度的位置局部性。表5显示了每个领域和层获得相同专家分配的连续token的比例，重复连续分配的比例显著高于随机分配。这对如何优化模型以实现快速训练和推理具有启示。例如，在执行专家并行时，具有较高局部性的情况更有可能导致某些专家的过度订阅。相反，这种局部性可以用于缓存。

**6 总结**

本文介绍了Mixtral 8x7B，这是第一个在开源模型中达到最先进性能的专家混合网络，在人类评估基准测试中胜过了Claude-2.1、Gemini Pro和GPT-3.5 Turbo。由于每个时间步只使用两个专家，Mixtral每个token只使用13B个活跃参数，而在性能上胜过了每个token使用70B参数的最佳模型（Llama 2 70B）。作者将训练和微调的模型公开发布，遵循Apache 2.0许可协议，旨在促进新技术和应用程序的发展，使各行业和领域受益。