原生多模态模型的缩放定律

Mustafa Shukor*
Sorbonne University

Enrico Fini
Apple

Victor Guilherme Turrisi da Costa
Apple

Matthieu Cord
Sorbonne University

Joshua Susskind
Apple

Alaaeldin El-Nouby
Apple

摘要

构建能够通过多模态信号有效感知世界的通用模型一直是一个长期目标。当前的方法涉及集成单独预训练的组件，例如将视觉编码器连接到大型语言模型（LLMs）并继续进行多模态训练。尽管这些方法展示了显著的样本效率，但是否这种后期融合架构本质上优于其他架构仍是一个开放问题。在本工作中，我们重新审视了原生多模态模型（NMMs）的架构设计——那些从底层开始就在所有模态上进行训练的模型——并进行了广泛的缩放定律研究，涵盖了 457 个具有不同架构和训练混合的模型。我们的研究表明，后期融合架构相对于早期融合架构没有内在优势，而早期融合架构不依赖于图像编码器。相反，早期融合架构在较低参数计数下表现出更强的性能，训练效率更高，部署更简单。受早期融合架构强大性能的启发，我们展示了一种引入混合专家模型（MoEs）的方法，使模型能够学习模态特定的权重，显著提升了性能。

1 引言

多模态为感知和理解世界提供了丰富的信号。视觉方面的进展，[????], 音频 [????] 和语言模型 [???] 使得开发能够理解语言、图像和音频的强大多模态模型成为可能。一种常见的方法是将分别预训练的单模态模型拼接在一起，例如，将视觉编码器连接到 LLM [??????????] 的输入层。

尽管这看起来是一种方便的方法，但这样的后期融合策略是否本质上是最优的 用于理解多模态信号 仍然是一个开放的问题。此外，随着大量多模态数据的可用性，从单模态预训练进行初始化可能会带来危害，因为它可能引入偏差，阻碍模型 充分利用跨模态的共依赖关系。另一个挑战是扩展此类系统；每个组件（例如视觉编码器、LLM）都有其自己的超参数集，预训练数据混合，以及 在数据和计算量方面 的扩展特性。一种更灵活的架构可能允许模型在模态之间动态分配其容量，从而简化扩展工作。

在这项工作中，我们关注从零开始在多模态数据上训练的原生多模态模型的缩放特性。首先，我们通过将常用的晚期融合架构与早期融合模型进行比较来研究它们是否具有内在优势，早期融合模型在处理原始多模态输入时不依赖专用视觉编码器。我们在早期和晚期融合

*在苹果公司实习期间完成的工作。

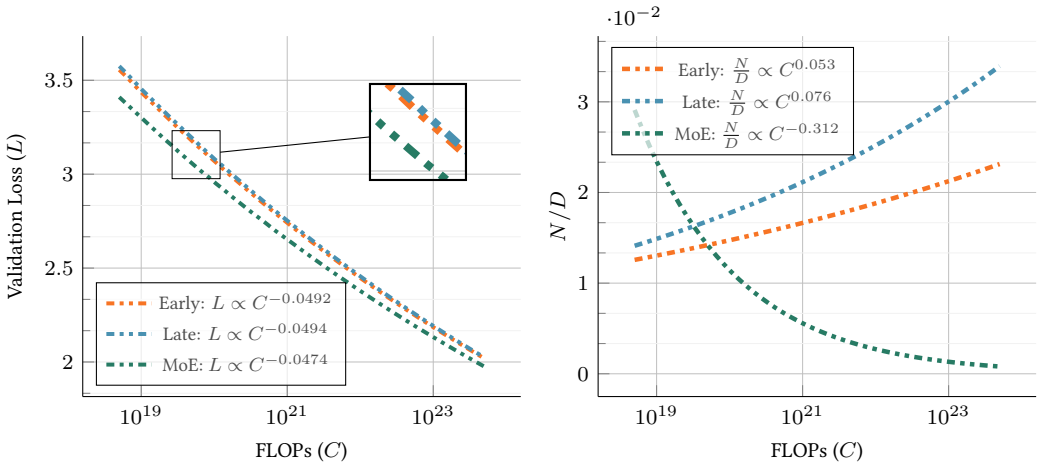


图 1: 原生多模态模型的缩放属性。基于在??中对缩放定律的研究，我们观察到：(1) 当使用相同的计算预算 C（以 FLOP 为单位）进行训练时，早期融合模型和晚期融合模型提供相当的验证损失 L；(2) 这种性能是通过在参数 N 和训练词元数量 D 之间进行不同的权衡来实现的，其中早期融合模型需要的参数更少。；(3) 对于给定的 FLOP 预算，稀疏早期融合模型实现了更低的损失并且需要更多的训练词元。

架构上进行缩放实验，推导出缩放定律以预测其性能和计算最优配置。我们的发现表明，当从头开始训练时，晚期融合没有固有的优势。相反，早期融合模型更高效且更容易扩展。此外，我们观察到原生多模态模型遵循与 LLMs 类似的缩放定律 [?]，尽管跨模态和数据集的缩放系数略有差异。我们的结果表明，为获得最佳性能，模型参数和训练词元应大致等比例缩放。此外，我们发现不同的多模态训练混合物表现出相似的整体趋势，这表明我们的发现很可能适用于更广泛的场景。

虽然我们的研究结果倾向于早期融合，但多模态数据本质上是异构的，这表明一定程度的参数特化可能仍然能带来好处。为 探索 这一点，我们 研究利用 混合专家（MoE）[?] 技术，该技术使模型能够以对称和平行的方式在模态间动态分配特化的参数，与晚期融合模型形成对比，后者是不对称的且按顺序处理数据。使用 MoE 训练原生多模态模型可以显著提高性能并 因此 加快收敛速度。我们的 MoE 扩展定律表明，扩展训练词元的数量比扩展活跃参数的数量更重要。这种不平衡扩展与密集模型所观察到的情况不同，因为稀疏模型具有更多的总参数。此外，我们的分析显示，专家倾向于在不同的模态上特化，这种特化在早期和最后一层尤为明显。

1.1 我们的发现摘要

我们的发现可概括如下：

原生早期融合和晚期融合表现相当： 从头开始训练的早期融合模型 与晚期融合对应模型表现相当，对于计算资源有限的情况，早期融合模型略有优势 (??)。此外，我们的缩放法则研究表明，随着计算预算的增加，早期和晚期融合的最佳计算模型表现相似 (?? 左侧)。

NMMs 缩放类似 LLMs： 原生多模态模型的缩放规律遵循与仅文本的 LLMs 相似的规律，缩放指数因目标数据类型和训练混合比例的不同而略有变化 (??)。

Expression	Definition
N	Number of parameters in the multimodal decoder. For MoEs this refers to the active parameters.
D	Total number of multimodal tokens.
N_v	Number of vision-only tokens.
D_v	Number of parameters in the vision-specific encoder. Only exists in late-fusion architectures.
C	Total number of FLOPs, estimated as $C = 6ND$ for early-fusion and $C = 6(N_v D_v + ND)$ for late-fusion.
L	Average validation loss on interleaved image-text, image-caption, and text-only data mixtures.

表 1: 论文中使用的表达方式的定义。

晚期融合需要更多的参数：与早期融合相比，计算最优的晚期融合模型需要更高的参数-to-数据比率（**??** 右）。

稀疏性显著有利于早期融合的 NMMs：在相同的推理代价下，稀疏 NMMs 相较于它们的稠密版本表现出显著的改进 (**??**)。此外，当使用稀疏性进行训练时，它们会隐式地学习模态特定的权重 (**??**)。此外，计算最优模型在计算预算增加时更多依赖于扩展训练词元的数量而不是活跃参数的数量 (**??** 右边)。

模态无关路由优于模态感知路由用于稀疏 NMMs：训练具有模态无关路由的稀疏混合专家模型始终优于具有模态感知路由的模型 (**??**)。

2 初步研究

2.1 定义

原生多模态模型 (NMMs): 从头开始同时在所有模态上训练的模型，不依赖于预训练的语言模型或视觉编码器。我们的重点是具有代表性的图像和文本模态，其中模型将文本和图像作为输入，并生成文本作为输出。

早期融合：从一开始就启用多模态交互，几乎不使用模态特定的参数（例如，除了将图像块片化的线性层之外）。使用单一的 Transformer 模型，该方法处理原始多模态输入——词元化的文本和连续的图像块，而无需对图像进行离散化。我们将主要的 Transformer 称为解码器。

晚融合：将多模态交互延迟到更深层，通常是在相互独立地处理每个模态的分离的单模态组件之后（例如，视觉编码器连接到大型语言模型）。

模态无关路由：在稀疏混合专家中，模态无关路由指的是依赖一个与模型联合训练的学成路由模块。

模态感知路由：基于预定义规则的路由，例如基于模态类型（如视觉-词元、词元-词元）的路由。

2.2 缩放律

我们旨在理解 NMMs 的缩放属性以及不同的架构选择如何影响权衡。为此，我们在 **??** 提出的缩放定律框架内分析我们的模型。我们基于总参数数量计算浮点运算次数（FLOPs），采用与先验工作 [**??**] 相同的近似公式 $C = 6ND$。然而，我们对这一估计进行了修改以适应我们

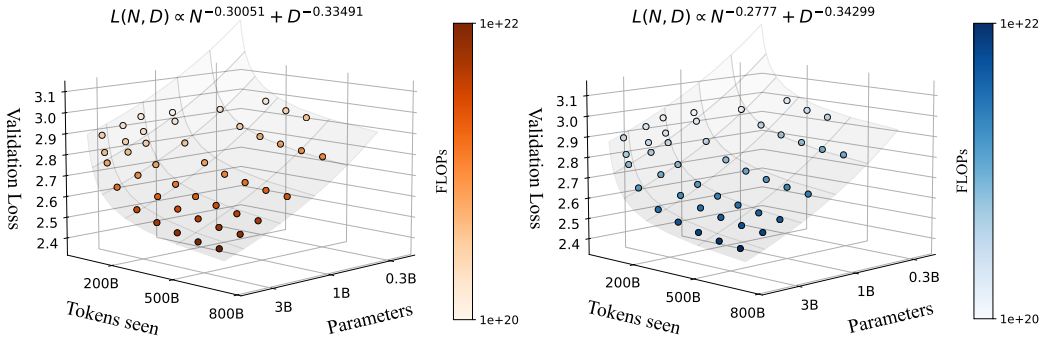


图 2: 适用于提前融合和延迟融合 原生多模态模型的缩放规律。每个点代表一个模型（300M 到 3B 参数），在不同数量的词元（250M 到 400B）上进行训练。我们在交错数据（Obelics）、图像-标题数据（HQITP）以及仅文本数据（DCLM）的有效验证集上报告平均交叉熵损失。

的设置：对于晚期融合模型，FLOPs 按照 $6(N_v D_v + N D)$ 进行计算。我们考虑一种设定，在给定计算预算 C 的情况下，目标是预测模型的最终损失，并确定最优的参数数量和训练词元数量。与 LLM 缩放相关的先前研究 [?] 一致，我们假设最终模型的损失与模型大小（N）和训练词元数量（D）之间存在幂律关系：

$$L = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}. \tag{1}$$

这里，E 表示在数据集上可实现的最低损失，而 $\frac{A}{N^{\alpha}}$ 捕获增加参数数量的效果，较大的模型会导致较低的损失，改进的速率由 α 控制。类似地，$\frac{B}{D^{\beta}}$ 考虑了更多词元数量的好处，β 确定改进的速率。此外，我们假设计算预算（FLOPs）与 N 和 D（$C \propto ND$）之间存在线性关系。这进一步导致了?? 中详细说明的幂律关系。

Data type	dataset	#samples	sampling prob.
	DFN [?]	2B	27%
Image-Caption	COYO [?]	600M	11.25%
	HQITP	400M	6.75%
Interleaved	Obelics [?]	141M Docs	45%
Text	DCLM [?]	6.6T Toks	10%

表 2: 预训练数据混合。除非另有说明，训练混合物包含 45%、45% 和 10% 的图像标题、交错文档和纯文本数据。

2.3 实验装置

Our models are based on the autoregressive transformer 架构 [?] 配备 SwiGLU 前馈网络 [?] 和 QK-归一化 [?] 后跟 ?。在早期融合模型中，图像块被线性投影以匹配文本词元的维度，而晚期融合则遵循 CLIP 架构 [?]。我们为文本词元采用因果注意力，为图像词元采用双向注意力，我们发现这种方法效果更好。训练在公共和私有多模态数据集的混合上进行，包括 DCLM [?]、Obelics [?]、DFN [?]、COYO [?] 和一个私有的高质量图文对 (HQITP) 集合（见 ??）。图像

被调整为 224×224 分辨率，使用 14×14 的块大小。我们对多模态序列使用 1k 的上下文长度。为了提高训练效率，我们在 bfloat16 中训练我们的模型，完全分片数据并行 (FSDP) [?]、激活 检查点和梯度累积。我们还对图像字幕数据集使用序列打包以减少填充词元的数量。类似于先前的工作 [???]，我们在交错（Obelics）、图像字幕（HQITP）和纯文本数据（DCLM）的保留子集上评估性能。进一步的实现细节见 **??**。

3 对原生多模态模型进行缩放

在本节中，我们研究了原生多模态模型的缩放规律，检查了各种架构选择 **??**，探索了不同的数据混合 **??**，分析了晚期融合和早期融合 NMMs 之间的实际权衡，并比较了原生预训练和连续预训练 NMMs 的性能 **??**。

设置。我们训练的模型从 0.3B 到 4B 个活跃参数，同时保持深度不变，仅缩放宽度。对于较小的训练词元预算，我们将预热阶段减少到 1K 步，而对于较大的预算则保持为 5K 步。遵循 **?**，模型以恒定学习率进行训练，随后使用逆平方根调度器进入冷却阶段。冷却阶段占总恒定学习率步骤的 20%。为了估计**??**中的缩放系数，我们应用了 L-BFGS 算法 [?] 和 Huber 损失 [?]（带 $\delta = 10^{-3}$），并在初始化值域范围内执行网格搜索。

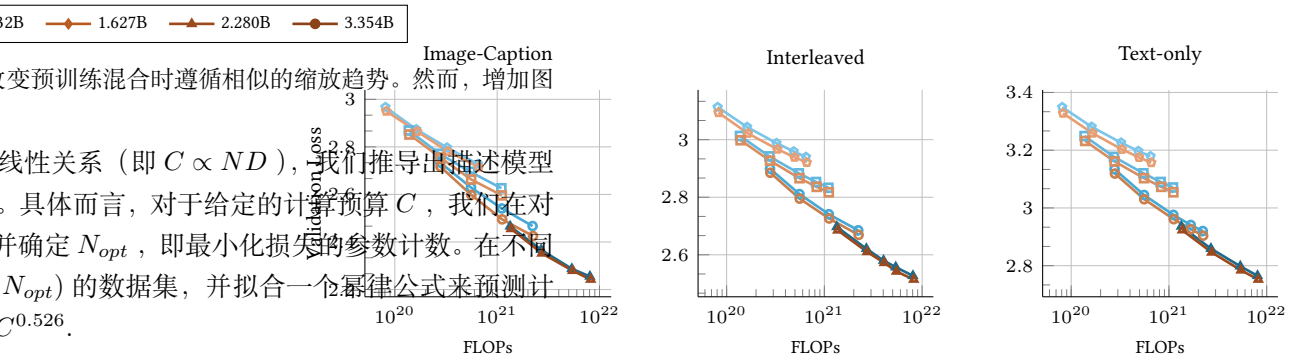
$L \propto E + \frac{1}{N^\alpha} + \frac{1}{D^\beta}$	$N \propto C^a$	$D \propto C^b$			$L \propto C^c$		$D \propto N^d$	
Model	Data	E	α	β	a	b	c	d
GPT3 [?]	Text	–	–	–	–	–	-0.048	
Chinchilla [?]	Text	1.693	0.339	0.285	0.46	0.54	–	
NMM (early-fusion)	Text	2.222	0.308	0.338	0.525	0.477	-0.042	0.909
	Image-Caption	1.569	0.311	0.339	0.520	0.479	-0.061	0.919
	Interleaved	1.966	0.297	0.338	0.532	0.468	-0.046	0.879
	AVG	1.904	0.301	0.335	0.526	0.473	-0.049	0.899
NMM (late-fusion)	AVG	1.891	0.290	0.338	0.636	0.462	-0.049	0.673
Sparse NMM (early-fusion)	AVG	2.158	0.710	0.372	0.361	0.656	-0.047	1.797

表 3: 原生多模态模型的缩放规律。我们报告了早期融合和晚期融合模型的缩放规律结果。我们为不同的目标数据类型及其平均损失（AVG）拟合了缩放规律。

3.1 NMMs 的缩放律

早融合和晚融合模型的缩放规律。 **??**（左）展示了早期融合 NMMs 在交错、图像-标题和文本数据集上的平均最终损失。最低损失前沿随着 FLOPs 的变化遵循幂律。拟合幂律得到表达式 $L \propto C^{-0.049}$，表明了计算增加时改进的速度。当按数据类型（如图像-标题、交错、文本）分析缩放规律时，我们观察到指数有所不同（**??**）。例如，与处理交错文档相比，该模型对图像-标题数据表现出更高的改进速度（$(L \propto C^{-0.061})$ 和 $(L \propto C^{-0.046})$。

将损失建模为训练词元数量 D 和模型参数 N 的函数，我们拟合了 **??** 中的参数化函数，得到缩放指数 $\alpha = 0.301$ 和 $\beta = 0.335$ 。它们分别描述了当扩展模型参数和训练词元的数量时改



变预训练混合时遵循相似的缩放趋势。然而，增加图

线性关系（即 $C \propto ND$），我们推导出描述模型
。具体而言，对于给定的计算预算 C，我们在对
并确定 N_{opt}，即最小化损失的参数计数。在不同
N_{opt}) 的数据集，并拟合一个幂律公式来预测计
$C^{0.526}$.

训练数据集大小与计算和模型大小的关系：

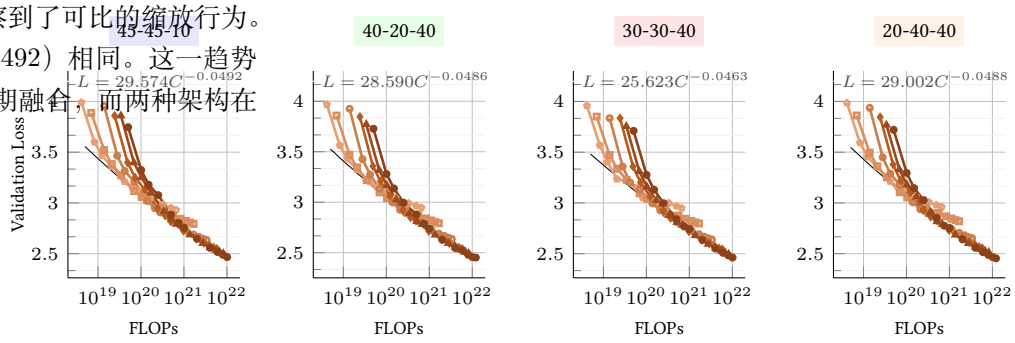
，　$D_{opt} \propto N^{0.899}$.

??

图 3: 早期融合与晚期融合：缩放训练浮点运算。我们比较了在缩放模型参数数量和训练词元数量时的早期融合模型和晚期融合模型。总体而言，早期融合显示出轻微的优势，尤其是在较小的模型规模下，随着参数数量的增加，差距减小 N 。

况下确定最佳模型和数据集大小。当按数据类型
2）相比图像-标题数据（$c = 0.520$）从更大的模
趋势。

了一项类似的研究，并观察到了可比的缩放行为。
乎与早期融合（$c = -0.0492$）相同。这一趋势
的模型规模下表现优于晚期融合，而两种架构在



1.965	0.328	0.348	0.518	0.486	0.937	-0.0486
1.847	0.253	0.338	0.572	0.428	0.748	-0.0463
1.836	0.259	0.354	0.582	0.423	0.726	-0.0488

还观察到，当改变晚期融合的配置时（例如，使
），也会出现类似的趋势。

我们 NMMs 的缩放律系数与报告中的文本专用
较，我们发现它们处于相似的范围内。特别是，在
循 $L \propto C^{-0.048}$ ，而我们的模型遵循 $L \propto C^{-0.049}$
的缩放规律。同样地，我们在 **??** （$\alpha = 0.301$ ，
告的值（$\alpha = 0.339$ ，$\beta = 0.285$ ）紧密匹配。同
3 值与 **?** 中的 $a = 0.46$ 和 $b = 0.54$ 高度一致，进
型，训练词元的数量和模型参数的数量应成比例
Ms 中更小，这一原则对 NMMs 更为适用。此外，
473 ，在给定计算预算的前提下，NMMs 的最佳
量则小于 LLMs。

NMMs）计算最优权衡。虽然晚期融合和早期融
损失，但我们观察到它们在计算最优模型中有不
模型更大，而 D_{opt} 对于早期融合模型更大。这表
模型需要更多的参数，而早期融合模型从更多的
早期融合较低的 $\frac{N_{opt}}{D_{opt}} \propto C^{0.053}$ 中，与晚期融合的
当扩展 FLOPs 时，早期融合模型的参数数量显著
务成本至关重要。

。我们
大的模型

b	0.35209	0.02892
a	0.54302	0.08813
c	0.43307	0.05787
d	1.92375	0.23296

表 6: Scaling laws sensitivity. We report the mean and standard deviation after bootstrapping with 100 iterations.

用??中的估计泛函形式计算损失，并将其与实际
这些比较，显示我们的估计在较低损失值和较大
估了我们的缩放律，预测了超出用于拟合的模型
方法合理地估计了 8B 模型的性能。

体来说，我们用带有替换的方式采样 P 个点（P
数。这个过程重复 100 次，我们报告每个系数的
在 α 上更精确，主要是由于用于推导缩放律的不
较少。

地多模态模型的缩放定律。为此，我们研究了四
[????], 其 Image Caption-Interleaved-Text 比例为
40-20-40 和 20-40-40 。对于每种混合物，我们按