少样本多模态多任务多语言学习

Aman Chadha计算机科学系 斯坦福大学
amanc@stanford.edu

Vinija Jain计算机科学系 斯坦福大学
vinija@stanford.edu

arXiv:2303.12489v1 [cs.LG] 19 Feb 2023

1 摘要

尽管作为一种迁移学习范式，少样本学习在数据有限的场景中已广受关注，但其研究主要集中于构建单模态、单语言模型。此外，现有少样本多任务学习领域的相当一部分研究采用上下文学习，需要将人工生成的提示作为输入，且效果会因人工提示工程的程度不同而有所差异。另外，上下文学习需要大量计算资源、内存和存储空间，导致推理延迟较高，因为每次进行预测时，都必须将提示中的所有示例输入模型。相比之下，基于微调范式进行迁移学习的方法，只需一次性按任务微调权重，即可避免上述问题。然而，这类方法尚未涉足少样本多模态多任务学习。本文提出面向多模态多任务多语言（FM3）场景的少样本学习方法：通过特定任务的超网络适配预训练视觉与语言模型，并对其进行对比式微调，从而实现少样本学习。FM3架构结合了上下文学习和基于微调的学习各自的优势，由三个主要部分组成：（i）通过多模态对比式微调实现少样本学习；（ii）通过超网络任务适配执行多任务学习；（iii）采用特定任务的输出头，以适应各种不同任务。FM3学习视觉与语言领域中最重要的任务及其交叉任务，包括视觉蕴含（VE） [1], 视觉问答（VQA） [2], 以及自然语言理解（NLU）任务，例如神经实体识别（NER）和GLUE基准[3] ，其中包括QNLI [4], MNLI [5], QQP [6],和SST-2 [7]。

2 引言

自监督预训练推动了深度学习在标注数据有限的任务中的应用。自监督预训练模型凭借任务无关特征和更高的数据效率，大幅降低了应对某些任务的机会成本；过去，这些任务需要大量数据，因此难以通过监督学习解决。随着自监督预训练不断发展，将自监督与针对相关任务的特定任务数据集上的监督学习相结合的半监督方法应运而生，成为一种促进迁移学习的新范式。

机器学习研究面临的最大开放挑战之一，是构建能够仅凭少量标注示例便快速适应新任务的模型。少样本学习（FSL）是迁移学习的一种特定变体，已成为解决标签稀缺场景的有效方案；在这些场景中，数据标注可能耗时且成本高昂。这些方法旨在利用少量带标签训练样例，通常涉及针对特定下游任务调整预训练模型。FSL方法有多种类型，各有优缺点。

随着生成式预训练变换器（GPT）系列 [8, 9] 自然语言处理模型的问世，一种大规模自监督方法广受欢迎：通过上下文学习（ICL）进行迁移学习，这种方法是在大规模训练中形成的。上下文学习通过向模型输入一个提示，其中包含少量监督示例，以及一个需要模型作出预测的未标注示例，从而教会模型执行下游任务。实际上，使用少量输入-目标对进行少样本提示，

可以为模型演示如何将输入转换为输出。值得注意的是，由于上下文学习无需参数更新，也就是说，不需要基于梯度的训练，因此单个模型能够立即执行各种各样的任务，实际上就像一把瑞士军刀。因此，上下文学习完全依赖于模型在预训练期间学到的能力。易于使用且能快速适应目标任务，是上下文学习得以广泛采用的典型特点 [10, 11, 12, 13, 14]。
尽管上下文学习（ICL）有诸多优点，但也存在几个主要缺点。首先，每次模型进行预测时，都要处理提示中的所有输入-目标对，这会带来巨大的计算资源、内存和延迟成本。随着推理次数增加，这些成本会不断累积——如果目标是在一批测试样例上执行推理，而不是只进行一次预测，那么从资源角度来看，上下文学习可能并不实用。其次，由于上下文窗口长度有限，模型可利用的支持样例 k 数量受到限制。这是因为我们必须将所有 k 示例放入模型的上下文窗口中，而该窗口能容纳的词元数量有限（GPT-2 为 1024 个，GPT-3 为 2048 个）。第三，与微调相比，上下文学习通常表现较差 [15, 8, 16]。最后，尽管模型的性能取决于提示的语义和结构方面，但这些方面可能对模型性能产生重大却难以预测的影响 [17], ，远远超出微调的运行间差异。具体而言，语义变化（例如提示的措辞或用词选择）以及句法变化（例如提示的具体格式，包括示例的措辞 [18] 和顺序 [19]）都可能对模型性能产生重大、非预期且难以估量的影响。此外，近期研究还表明，即使提供错误标签，上下文学习也能取得良好表现，这引发了人们对其究竟在多大程度上发生了学习的疑问 [13]。
另一种常见的半监督学习范式是通过微调（FT）进行迁移学习，它遵循两阶段流程：（i）利用预训练的大规模自监督模型的参数进行权重初始化；（ii）使用与目标下游任务相关的数据执行基于梯度的微调。随着 BERT 等表征学习方法的出现， [20], 自然语言处理领域发生了彻底转变：对于情感分析、神经实体识别、问答、摘要生成、对话响应生成等任务，研究范式从监督方法转向了半监督方法。如今，表征学习方法已成为自然语言处理的核心；这些预训练模型学习到的上下文化表征可作为初始的任务无关特征，进而为学习任务特定特征提供起点。尽管标注数据有限的问题因半监督方法对数据的需求较低而获益显著，但标注数据丰富的任务也取得了更好的性能。

尽管微调已在一系列分类任务上取得了许多最先进水平的结果， [21] 但它会生成一个专门针对单一任务、参数值全新的一组模型；当需要对模型进行微调以适配许多下游任务时，这可能不切实际。换言之，这类模型通常一次只能执行一项任务，也无法通过少量样本学习新概念或适应新任务。FM3 旨在弥补这一不足，实现多模态少样本学习——与 SETFIT 的做法类似：它通过对比学习微调预训练的 Sentence Transformer 模型， [22]完全摒弃提示，并且无需大规模预训练语言模型即可达到很高的准确率。在 Customer Reviews（CR）情感数据集中仅使用 8 个带标签样本时，SETFIT 的表现足以媲美使用完整训练集微调的 RoBERTa，[23],尽管后者的模型规模是其三倍。

上下文学习和微调都已在多模态场景中得到探索。许多方法，尤其是*Flamingo* [15] 和*Frozen* [24], ，采用上下文学习，最终目标是使模型能够迅速适应各种多模态任务。虽然*Flamingo* 取得了与少样本学习相当的性能，在某些情况下甚至优于使用多出数千倍特定任务数据进行微调的模型，但*Frozen* 的性能相对较低，换来的是能够灵活使用现成的预训练语言模型，并保持其权重冻结。另一方面，
Oscar [25] 和 Omninet [26] 是不会进行上下文学习的多模态多任务模型。Oscar 使用对齐数据，在任务无关的跨模态目标（对词语和视觉标签计算掩码词元损失，以及计算视觉标签与其他内容之间的对比损失）上进行预训练，然后针对特定任务进行微调；而 Omninet 则同时在其目标任务上接受训练，且不进行微调。在零样本／少样本学习场景中，近期研究表明，多模态预训练能够通过大规模对比学习，在判别式设置下实现强大的泛化能力 [27, 28]。

让模型以最少的更新来执行新任务的另一种范式是参数高效微调（PEFT）：通过仅更新少量新增或选定的参数，对预训练模型进行微调。近期方法仅更新或添加

模型全部参数中的一小部分（例如 0.01%），便达到了与对整个模型进行微调相当的性能 [29, 30]。此外，某些参数高效微调方法支持混合任务批次，其中同一批次中的不同示例会以不同方式处理 [30], ，因此参数高效微调和上下文学习都适用于多任务模型。尽管参数高效微调的优势弥补了微调（相较于上下文学习）的一些不足，但对于标注数据极少时参数高效微调方法是否有效，人们关注得相对较少。 [16]通过提出 T-Few 填补了这一空白。T-Few 是一种采用参数高效微调和固定超参数进行学习的模型，在新颖、未见过的任务上表现出色，同时仅更新模型中极小一部分参数。

FM3 融合了基于 ICL 和 FT 的迁移学习各自的优势，提供了一种高效、无提示的框架，能够在少样本设置下对新的多模态视觉语言任务实现强泛化。尽管 ICL 灵活性很高，但它的局限性仍有待改善，尤其是在计算资源、延迟、内存、批量推理、性能确定性等因素至关重要的情况下。另一方面，FT 无需提示，因此能够保持性能稳定；其性能优于基于 ICL 的方法 [15],，并且在计算资源、延迟、内存等方面更加高效。虽然零样本/少样本泛化是 ICL 带来的一个理想副产品，但 FT 唯一显著的缺点是，在数据有限时，很难泛化到新任务。FM3 在设计时考虑了上述 ICL 的缺陷，因此采用 FT 方法，同时通过以下方式克服其局限性：(i) 使用多模态对比式微调实现少样本学习；(ii) 使用参数量有限的超网络执行多任务学习的任务适配；(iii) 采用任务特定输出头，以适应众多不同任务。

FM3 仅需少量标注数据即可实现高准确率——例如，在 SNLI-VE 这一复杂任务中，每个类别只使用 16 个带标签样本，FM3 的表现便超越了当前最先进水平：后者是在包含 43 万个示例的完整训练集上微调得到的！与其他 FSL方法相比，FM3 具有以下几项独特优势：

- **无需提示或 verbalizer：** 当前的少样本微调技术需要手工设计提示或 verbalizer，将示例转换为适合底层语言模型的格式。FM3 完全不需要提示，而是直接从文本示例生成丰富的嵌入。这彻底消除了手动进行提示工程的必要，进而确保了性能确定性。

- **资源效率：** 与我们的基线*Flamingo*、*Frozen*，尤其是基于 ICL 的方法相比，FM3 能够更优地利用计算资源，并在延迟、内存等方面表现出色。
- **冻结的预训练模型：** FM3 使用预训练的视觉和语言编码器，而不对其进行微调。这意味着 FM3 架构支持直接即插即用地替换模态编码器。使用不同编码器进行实验时，只需微调小型超网络模型。
- **训练速度快：**FM3 无需使用*Flamingo*（80B）或*Frozen*（7B+）这类大型模型，也能达到较高的准确率。因此，训练和推理速度都显著更快。
- **支持多语言：**FM3 支持多语言处理，并且可以与任何多语言文本编码器配合使用，例如多语言 Sentence Transformer [22] 、MPNet [31],、RoBERTa [32], 、ALBERT [33], 、LASER [34], 等的变体。只需微调一个多语言模型检查点，即可支持 50+ 种语言的多语言学习。

虽然已有研究提出了针对少样本多模态多任务多语言学习部分领域的方法，但据我们所知，FM3 是首个在少样本学习设置下探索多模态多任务多语言学习各领域交叉的研究。

3 相关工作

3.1 使用预训练模型进行少样本学习

在自然语言处理领域，Tunstall 等人提出的 SETFIT [23]是一个高效且无提示的框架，用于对 Sentence Transformers（ST）进行少样本微调。SETFIT 通过以对比式孪生网络方式，在少量文本对上微调预训练 ST 来实现这一点。随后，使用所得模型生成丰富的文本嵌入，并以此训练分类头。这个简单的框架无需提示，与现有技术相比，所需参数量少了几个数量级，同时仍能达到很高的准确率。SETFIT 的效果可与参数高效微调（PEFT）和参数高效调优（PET）技术相媲美，而训练速度快一个数量级。SETFIT 只需少量标注数据即可取得很高的准确率——例如，在客户评论

情感数据集上，每个类别仅使用 8 个带标签样本，SETFIT [35], 便能与在包含 3K 个示例的完整训练集上微调 RoBERTaLarge 的效果相媲美。鉴于其在实现少样本学习方面的实用价值，我们借鉴 SETFIT 中对比式微调的思路，并将其推广至多模态多任务多语言设置，作为 FM3 的一部分。

3.2 使用参数高效微调进行多任务微调

在 [36], Houlsby 等人提出了一种参数高效微调方法，该方法在预训练语言模型的各层之间引入适配器模块。适配器模块构成紧凑且易于扩展的模型；每个任务仅需增加少量可训练参数，而且可以添加新任务而无需重新处理先前的任务。原始网络的参数保持固定，从而实现高度的参数共享。他们在 GLUE [3] 基准上取得了最先进水平，同时每个任务仅增加少量参数。不过，这种方法的缺点是每个任务都要单独训练，因此无法实现跨任务的信息共享。

为解决上述跨任务知识共享问题，Mahabadi 等人 [37] 使用共享超网络为所有层和任务学习适配器参数；这些超网络以任务、适配器位置和 Transformer 模型中的层 ID 为条件。该参数高效多任务学习框架兼顾两方面优势：通过超网络共享跨任务知识，同时借助任务专用适配器使模型能够适应各个独立任务。在 GLUE基准上进行的实验表明，该方法提升了多任务学习的性能，同时每个任务仅增加 0.29% 的参数。鉴于超网络无需实际微调模型权重（也就是说，这些权重在此过程中保持 Frozen），便能轻松实现预训练模型的多任务微调，因此我们在所提出的架构中采用多任务微调。

3.3 多任务多模态学习

Hu 和 Singh 提出了 UniT [38], 统一 Transformer 编码器-解码器模型，该模型在分布于不同视觉和语言领域的 8 个数据集上联合学习 7 项任务，涵盖从目标检测到自然语言理解和多模态推理等任务。UniT 仅使用少量参数便取得了出色的性能，在某些情况下甚至优于分别训练的单任务模型。尽管该架构支持对每项任务进行联合端到端训练，但要使模型具备泛化能力，就需要所有任务的大量数据。相比之下，我们的方法利用少样本学习，只需少量数据便能高效地学习一项任务。

在 [26], Pramanik 等人提出了 OmniNet，这是一种支持多种输入模态任务以及异步多任务学习的单一模型。OmniNet 由时空缓存驱动，除了与时间输入序列相对应的隐藏状态外，还能学习输入的空间维度。尽管 OmniNet 的参数效率较高，但与各项任务对应的单独模型相比，它在大多数训练过的任务上仍存在明显的性能差距。

3.4 多任务多语言多模态学习

M^3P在 [39], 提出，是一种多任务多语言多模态预训练模型，通过多任务预训练将多语言预训练和多模态预训练整合到统一框架中。 M^3P学习通用表征，可将不同模态中的对象或以不同语言表达的文本映射到共同语义空间。此外，为缓解非英语多模态任务缺乏充足标注数据的问题，他们提出了多模态代码转换训练（MCT）[40]： 以概率 β将图注中的每个词替换为其译词。如果一个词有多个译法，则随机选择一个。在 MS COCO [41] 和 Multi30K [42] 上进行的多语言图像检索任务实验表明，该方法在英语任务上取得了有竞争力的结果，并在非英语语言任务上创下新的最先进水平。虽然 M^3P 解决的问题与 FM3 类似，但它不对标注预算做任何限制；换言之，它并未考虑以少样本设置来学习其任务。

3.5 少样本多模态多任务学习

在 [15], Alayrac 等人提出了 Flamingo，这是一个视觉语言模型（VLM）系列，基于大规模多模态网络语料库训练，能够快速适应各种图像和视频任务。*Flamingo* 提出了以下关键架构创新：(i)连接强大的预训练纯视觉模型和纯语言模型；(ii)处理视觉和文本数据任意交错的序列；并且

(iii)无缝接收图像或视频作为输入。最终得到的是一个统一的 *Flamingo* 模型，只需通过特定任务示例提示模型，便能凭借少样本学习达到新的最先进水平。在众多基准测试中，*Flamingo* 的表现优于那些使用多出数千倍的特定任务数据进行微调的模型。这些任务包括开放式任务，例如视觉问答，模型会收到问题提示并需要作答；图像描述任务，用于评估模型描述场景或事件的能力；以及选择题式视觉问答等封闭式任务。

在 [24], Tsimpoukelli 等人介绍了*Frozen*，这是一种简单而有效的方法，可将大型自回归语言模型固有的少样本学习能力迁移到多模态设置（视觉和语言）中。*Frozen*是一种多模态少样本学习器，具有令人惊讶的能力：在以示例为条件时，能够学习各种新任务，这些示例以多个交错排列的图像和文本嵌入序列表示。*Frozen* 只需少量示例，就能迅速学会表示新物体和新颖视觉类别的词语，并完成视觉问答。尽管这项工作为 FM3 提供了重要的基线，但其关键局限在于，在少样本设置下，它在所学习的特定任务上的表现远未达到最先进水平。*Frozen* 表明，通过预训练且冻结的语言模型训练视觉编码器，可以构建出具备强大分布外（零样本）泛化能力的系统。此外，*Frozen* 证实，在获得恰当提示时快速适应新任务的能力源自预训练语言模型，并可直接迁移到多模态任务中。

虽然 *Flamingo* 和 *Frozen*都是基于上下文学习的少样本学习方法，但二者的区别在于：（i）用于训练这些模型的数据规模；（ii）架构上的差异。*Flamingo* 在大规模多模态网络语料库上训练，而*Frozen* 则在 Conceptual Captions 数据集 [43]上训练。两者在架构设计上的选择不同：一方使用预训练的模态编码器，另一方则从头开始训练。与 FM3 类似，*Flamingo* 使用现成的预训练编码器，仅生成适配器组件（以 Perceiver Resampler 模块的形式）；而*Frozen* 使用预训练语言模型，但会训练自己的视觉编码器并将其输出馈送给语言模型。受这一观察启发，FM3 借鉴了这一思路，即

使用形式为超网络的独立文本适配器和视觉适配器，为模型提供额外的自由度，从而有助于提升模型性能。

4 FM3

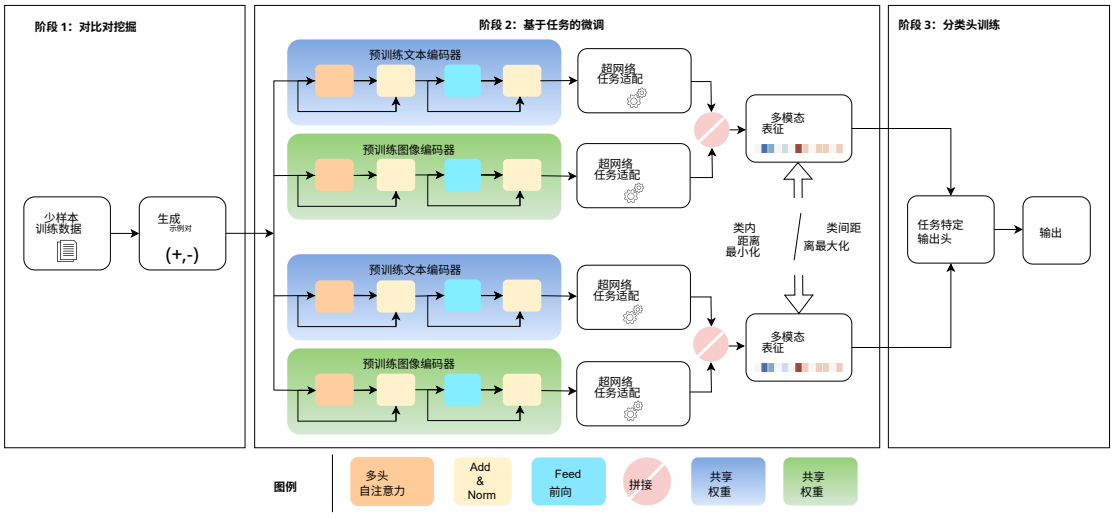


图 1：FM3 的架构概览。FM3 由三个阶段组成：(i) 用于微调的对比配对挖掘，用于生成正样本对和负样本对；(ii) 基于任务的微调，通过超网络调整预训练文本和图像编码器模型，以适应下游任务；(iii) 训练任务特定分类头。

4.1 方法

图 1 直观总结了 FM3 的各个架构阶段。

4.1.1 任务与批次采样

在训练期间的每次迭代中，我们都会随机选择一个任务，其采样概率可根据数据集大小手动指定。任务列表采样完成后，对于包含多个数据集的任务，我们会随机采样一个与该任务对应的数据集，以填充一个样本批次。

4.1.2 用于少样本学习的对比式微调

与 [23], 类似，我们采用对比学习方法进行少样本学习。对比学习能够有效扩大训练数据的规模，这在小样本场景中至关重要，因此有助于在标注数据有限的情况下有效学习任务。假设某分类任务有少量（k）带标签样本，那么根据可生成的唯一正负对比样本对数量推导出的微调集潜在规模 T 将达到 $^{k}_{(k-1)\ 2}$，远大于仅有 k [23] 的规模。在此阶段，我们采样 R 个正三元组对和 R 个负三元组对，其中 R 是一个超参数（遵循 [23]，设为 20）。鉴于多负例排序损失 [44] 性能更优[44]，且能够自动从每个批次中随机采样负样本对，我们采用该损失进行对比式微调。

4.1.3 使用超网络进行基于任务的微调

据我们所知，FM3 是首个在多模态设置中使用超网络的方法。使用冻结的模态编码器具有一项显著优势：可以防止灾难性遗忘（与微调编码器本身相比） [15]。因此，我们为每种模态分别使用一个独立的超网络。

在此步骤中，我们对最先进的预训练文本与视觉模型进行任务专用微调，具体而言，使用预训练的多语言 MPNet [31] Sentence Transformer [22] from Huggingface [45] 作为文本骨干网络，并使用 CoCa-Base [46] 作为视觉骨干网络。我们借鉴了 [37],中的超网络理念，这是一种用于多任务微调的参数高效方法。我们训练共享超网络，根据任务、层 ID 和适配器位置嵌入生成任务专用适配器。这些共享超网络能够捕获跨任务知识，从而促进向低资源及相关任务的正迁移；而任务特定层则使模型能够适应各个具体任务。我们优化一个基于余弦相似度的距离函数：对正样本对最小化该函数，对负样本对最大化该函数。

4.1.4 任务特定分类头训练

最后，我们在上一步获得的微调模型上训练任务特定分类头。每项任务的数据样本所对应的生成嵌入及其标签，构成相应分类头的训练集。对于 SST-2、QQP、QNLI 等二分类任务，我们使用逻辑回归；对于 VQA、SNLI-VE、GLUE、NER 等多分类任务，我们使用 Softmax。

4.2 任务和数据集

表1列出了 FM3 训练和评估所涉及的领域、任务和数据集。

领域	Task	数据集
语言理解	神经实体识别（NER）	CoNLL-2003 [47]
	GLUE基准 [3]	QNLI [4], MNLI [5], QQP [6],和 SST-2 [7]
视觉与语言推理	视觉蕴含	SNLI-VE [1]
	视觉问答（VQA）	VQAv2 数据集 [2]（以 Visual Genome [48] 中的问题作为额外数据），OK-VQA [49]

表1：用于训练和评估的数据集

5 实验

5.1 实验设置

我们在 Conceptual Captions 数据集（*Frozen* [24] 正是在该数据集上训练的）上对 FM3 进行了微调，以便进行直接比较。我们使用 AdamW 优化器，并进行全局范数裁剪，阈值为 1；超网络不使用权重衰减，其他可训练参数的权重衰减设为 0.1。我们采用退火学习率：在最初 5000 步中将学习率从 0 线性提升至 10^{-3}，之后在训练期间保持不变，随后再按指数衰减。除非另有说明，我们的模型均训练 500K 步。

所有数据集均使用相同的权重进行训练。由于使用对比目标训练的模型性能对批量大小较为敏感，我们采用了相对较大的批量大小 32。

5.2 基线

我们采用 *Flamingo* [15] 和 *Frozen* [24] 作为主要基线，因为它们处理的是少样本学习背景下的多模态多任务学习。我们还将 UniT [38] 作为补充基线，因为它处理的是 *Flamingo* [15]和 *Frozen* 尚未评估过的任务的多模态多任务学习。对于每项任务，我们将 FM3 与任务专属的零样本/少样本方法以及预训练/微调的最先进水平进行比较。由于上述基线均不支持多语言任务，我们采用 [39] 作为基线，以评估 FM3 在非英语任务上的表现。

5.3 结果

方法	FT	样本数	OKVQA [49]	VQAv2 [2]	Flickr30K [50]	Multi30K [42]	SNLI-VE [1]	CoNLL-2003 [47]	QNLI [4]	MNLI [5]	QQP [6]	SST-2 [7]
零样本/少样本少样本最先进水平	✗	各种	[51] 43.3 (16)	[24] 38.2 (4)	[52] 94.9 (0)	-	[53] 87.3 (0)	[54] 65.4 (5)	-	[53] 86.4 (5)	[55] 67.8 (16)	[55] 93.5 (16)
Flamingo-80B	✗	0	50.6	56.3	67.2							
		4	57.4	63.1	75.1	-	-	-	-	-	-	-
		32	57.8	67.6	75.4							
Frozen	✗	0	5.9	29.5								
		1	9.7	35.7	-	-	-	-	-	-	-	-
		4	12.6	38.2								
UniT	✗	完全有监督	-	67.0	-	-	73.2	-	88.0	81.8	90.6	91.5
FM3-5B	✗	0	51.5	57.3	93.2	27.2	86.2	51.3	53.6	58.2	41.4	75.5
		4	55.5	64.6	94.3	35.2	89.9	66.2	67.6	55.5	54.4	86.8
		16	57.8	67.9	96.1	37.1	92.4	77.3	75.4	81.1	66.7	95.7
		64	**58.9**	**71.2**	**99.1**	39.9	**94.4**	**82.7**	83.7	86.4	**78.9**	**99.2**
预训练微调 最先进水平	✔	各种	54.4 [51] (10K)	80.2 [56] (444K)	98.8 [57] (31K)	49.3 [58] (31K)	91.0 [59] (430K)	94.6 [60] (430K)	99.2 [33] (105K)	92.0 [61] (393K)	89.2 [62] (364K)	97.5 [61] (67K)

表2：**与最先进水平的比较**。 单个 FM3 模型在广泛的视觉-语言理解任务上通过少样本学习达到最先进水平，仅用 16 个示例就显著优于以往最佳的零样本和少样本方法。更重要的是，只使用 64 个示例且不调整任何模型权重，FM3 在 4 项任务上超越了当前最佳方法——这些方法均使用数千个标注示例进行微调。所有样本数下最佳的少样本数值以**粗体**显示；零样本/少样本（基于提示）和微调模型中的最佳数值均加下划线。对于每个基线，我们选取了涵盖所有变体/实验的最佳数值（除非另有明确说明）。

表2对 FM3、*Flamingo*、*Frozen*以及各任务上相应的零样本/少样本和微调最先进水平进行了比较，支持示例/样本数为 $k \in \{0, 4, 16, 64\}$。Flickr30K 图像到文本使用 Recall@1，Mulit30K 英德使用 BLEU，CoNLL-2003 和 QQP 使用 F1，其他所有任务均以准确率作为性能指标。请注意，由于 *Frozen*是一个采用基于提示的微调的自回归模型/解码器， k 表示作为提示/前缀的一部分传递给模型作为输入的支持示例数量；而 FM3 是基于编码器的架构， k表示进行对比微调所用的示例数量。

少样本结果。 在所考察的 10 个基准测试中，FM3 有 7 个优于零样本／少样本基线。每项任务只需 16 个示例即可实现这一成果，展现了对这些任务的出色适应能力。更重要的是，FM3 的表现往往可与使用多达数十万个标注示例进行微调的最先进方法相媲美。在 10 项任务中，有 4 项即使只使用一组模型权重和 64 个特定任务示例，FM3 的表现也优于经过微调的最先进方法。
参数和样本数的扩展。如表2所示，样本数越多，少样本性能越好，这与 GPT-3 [8] 类似。随着样本数增加，性能提升呈现边际收益递减。

5.4 推理运行时分析

表3总结了我们的推理运行时间分析。我们测量了 FM3 和主要基线在 VQAv2 [2] 和 OKVQA [49] 测试集上运行所需的时间，并除以总样本数取平均值。这些测量是在配备 32GB 显存的 NVIDIA A100 平台上完成的。**粗体**数字表示最佳性能。带下划线的数字表示次优基线，FM3 的百分比提升以此为依据。

	VQAv2（秒）	**OKVQA**（秒）
FM3	**0.187** (58% ↑)	**0.214** (46% ↑)
Flamingo	0.353	0.371
Frozen	0.295	0.312

表3：FM3 与*Flamingo*和 *Frozen* 在 VQAv2 和 OKVQA 上的推理运行时间分析。测量结果基于墙上时钟时间（秒），因此越低越好。

6 消融分析

表4总结了 FM 的消融实验结果。**粗体**数字表示最佳性能。带下划线的数字表示作为百分比计算基准的基线。我们分析以下设计决策对 FM3 性能的影响：

- **不使用超网络，直接微调编码器**。虽然冻结模态编码器可以防止灾难性遗忘， [15], 但我们采用无超网络架构，并使用目标任务的数据直接微调编码器。
- **超网络规模选择**。我们比较了超网络的不同参数量配置，以量化超网络规模的影响。超网络的参数量以其占 FM3 基线模型参数量的百分比表示。
- **计算资源/内存与性能之间的权衡**。 我们改变文本和视觉编码器的选择（这也会改变模型的参数数量和时间复杂度）。对于文本编码器，我们选择多语言 MiniLM [63] ，其参数量比默认选择的多语言 MPNet 的 paraphrase-multilingual-mpnet-base-v2 变体少 57% [45]。对于视觉编码器，我们选择 ViT[64]的 vit-base-patch16-224 变体，其参数量比默认选择的 CoCa-Base 少 78% [46]。

	VQAv2（准确率）	**OKVQA**（准确率）
FM3（默认）	**71.2**	**58.9**
无超网络	64.3（90.0% ↓）	52.1（88.4% ↓）
使用 5% 参数的超网络	69.5（97.6% ↓）	56.7（88.4% ↓）
文本编码器：MiniLM	68.1（95.6% ↓）	55.4（94.0% ↓）
视觉编码器：ViT	66.2（92.9% ↓）	51.2（86.9% ↓）
文本/视觉编码器：MiniLM/ViT	65.5（91.9% ↓）	52.3（88.7% ↓）

表4：在样本数为64的设置下，FM3在VQAv2和OKVQA上的消融分析。默认的FM3使用参数量为10%的超网络，并采用MPNet/CoCa作为文本/视觉编码器。测量单位为%准确率，数值越高越好。

7 未来工作

尽管 FM3 在多项任务上达到了新的最先进水平，但仍有很大的改进空间，主要集中在三个方面：（i）数据；（ii）模型架构；（iii）损失函数。首先，*Flamingo*

[15] 强调了在神经网络训练中，汇集各种不同来源的数据集的重要性（*Flamingo*使用了 >2B 个图文对，而 FM3 的训练数据为 330 万个）。使用公开的大规模 LAION-400M 数据集 [65] 将是一个很好的起点。其次，模型架构可以采用其他技术，在减少参数量的同时实现相当高的性能，例如基于低秩的适配方法 LoRA [29]。第三，按照 [66, 25],的做法，我们可以将排序损失 [44] 表述为一个二分类问题。据报道，这样做能够提升性能 [66, 25]。换言之，给定一个对齐的图文对，我们随机选取另一张图像或另一条图注，组成一个未对齐的示例对。与 FM3 当前的框架类似，最终拼接得到的多模态嵌入仍可作为分类的输入，用于预测给定的示例对是否对齐。最后，FM3 可以轻松扩展到其他语言、任务和模态。

8 结论

FM3 融合了上下文学习与微调各自的优势，成为少样本多语言多模态多任务学习这一细分领域的领先者。它提供了一种可扩展的架构，能够涵盖多种模态、任务和语言，同时在众多任务上达到最先进水平，并在其他任务上展现出有竞争力的表现。每项任务仅需 16 个示例，FM3 在 10 个基准测试中的 7 个上就优于零样本／少样本基线。此外，在使用多达数十万个标注示例进行微调的任务专用最先进模型中，FM3 的表现也具有竞争力。在 10 项任务中，有 4 项中 FM3 甚至优于微调后的最先进模型，同时仅使用一套模型权重和 64 个特定任务示例。最后，在 VQA 和 OKVQA 数据集上，与表现次佳的少样本学习最先进基线相比，FM3 的延迟还降低了 ∼50%。

参考文献

[1] Ning Xie、Farley Lai、Derek Doran 和 Asim Kadav。视觉蕴含：一项用于细粒度图像理解的新任务。*arXiv* 预印本 *arXiv:1901.06706*，2019年。[2] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。让 VQA 中的“V”发挥作用：提升图像理解在视觉问答中的作用。载于*IEEE*计算机视觉与模式识别会议论文集，第 6904–6913 页，2017年。[3] Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R Bowman。GLUE：自然语言理解的多任务基准与分析平台。*arXiv* 预印本 *arXiv:1804.07461*，2018年。[4] Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev 和 Percy Liang。Squad：用于文本机器阅读理解的 100,000+ 个问题。*arXiv* 预印本 *arXiv:1606.05250*，2016年。[5] Adina Williams、Nikita Nangia 和 Samuel R Bowman。通过推理实现句子理解的广覆盖挑战语料库。*arXiv* 预印本 *arXiv:1704.05426*，2017年。[6] Adina Williams、Nikita Nangia 和 Samuel R Bowman。首个 Quora 数据集发布：问题对。*Quora* 博客，2017年。[7] Richard Socher、Alex Perelygin、Jean Wu、Jason Chuang、Christopher D Manning、Andrew Y Ng 和 Christopher Potts。情感树库上的语义组合性递归深度模型。载于*2013*年自然语言处理实证方法会议论文集，第 1631–1642 页，2013年。[8] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。语言模型是无监督多任务学习器。*OpenAI* 博客，1(8):9，2019年。[9] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习器。神经信息处理系统进展，33:1877–1901，2020。[10] Sewon Min、Mike Lewis、Luke Zettlemoyer 和 Hannaneh Hajishirzi。MetaICL：学习上下文中的学习。*arXiv* 预印本 *arXiv:2110.15943*，2021年。[11] Andrew K Lampinen、Ishita Dasgupta、Stephanie CY Chan、Kory Matthewson、Michael Henry Tessler、Antonia Creswell、James L McClelland、Jane X Wang 和 Felix Hill。语言模型能否从上下文中的解释中学习？*arXiv* 预印本 *arXiv:2204.02329*，2022年。

[12] Angeliki Lazaridou、Elena Gribovskaya、Wojciech Stokowiec 和 Nikolai Grigorev。通过少样本提示增强互联网语言模型，以回答开放域问题。*arXiv* 预印本 *arXiv:2203.05115*，2022年。[13] Sewon Min、Xinxi Lyu、Ari Holtzman、Mikel Artetxe、Mike Lewis、Hannaneh Hajishirzi 和 Luke Zettlemoyer。重新思考示例的作用：上下文学习为何有效？*arXiv* 预印本 *arXiv:2202.12837*，2022年。[14] Yizhong Wang、Swaroop Mishra、Pegah Alipoormolabashi、Yeganeh Kordi、Amirreza Mirzaei、Anjana Arunkumar、Arjun Ashok、Arut Selvan Dhanasekaran、Atharva Naik、David Stap 等。通过上下文指令对 1,600+ 个语言任务的泛化能力进行基准测试。*arXiv* 预印本 *arXiv:2204.07705*，2022年。[15] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katie Millican、Malcolm Reynolds 等。Flamingo：用于少样本学习的视觉语言模型。*arXiv* 预印本 *arXiv:2204.14198*，2022年。[16] Haokun Liu、Derek Tam、Mohammed Muqeeth、Jay Mohta、Tenghao Huang、Mohit Bansal 和 Colin Raffel。少样本参数高效微调优于上下文学习，且成本更低。*arXiv* 预印本 *arXiv:2205.05638*，2022年。[17] Yao Lu、Max Bartolo、Alastair Moore、Sebastian Riedel 和 Pontus Stenetorp。排列精妙的提示及其寻觅之道：克服少样本提示顺序敏感性。*arXiv* 预印本 *arXiv:2104.08786*，2021年。[18] Albert Webson 和 Ellie Pavlick。基于提示的模型真的理解其提示的含义吗？*arXiv* 预印本 *arXiv:2109.01247*，2021年。[19] Zihao Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。使用前先校准：提升语言模型的少样本性能。载于国际机器学习会议，第 12697–12706 页。PMLR，2021年。[20] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018年。[21] Victor Sanh、Albert Webson、Colin Raffel、Stephen H Bach、Lintang Sutawika、Zaid Alyafeai、Antoine Chaffin、Arnaud Stiegler、Teven Le Scao、Arun Raja 等。多任务提示训练实现零样本任务泛化。*arXiv* 预印本 *arXiv:2110.08207*，2021年。[22] Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。载于*2019*年自然语言处理实证方法会议论文集。计算语言学协会，2019年11月。[23] Lewis Tunstall、Nils Reimers、Unso Eun Seo Jo、Luke Bates、Daniel Korat、Moshe Wasserblat 和 Oren Pereg。无需提示的高效少样本学习。*arXiv* 预印本 *arXiv:2209.11055*，2022年。[24] Maria Tsimpoukelli、Jacob L Menick、Serkan Cabi、SM Eslami、Oriol Vinyals 和 Felix Hill。使用冻结语言模型进行多模态少样本学习。神经信息处理系统进展，34:200–212，2021年。[25] Xiujun Li、Xi Yin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei 等。Oscar：面向视觉语言任务的物体语义对齐预训练。载于欧洲计算机视觉会议，第 121–137 页。施普林格，2020。[26] Subhojeet Pramanik、Priyanka Agrawal 和 Aman Hussain。OmniNet：用于多模态多任务学习的统一架构。*arXiv* 预印本 *arXiv:1907.07804*，2019年。[27] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。从自然语言监督中学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021年。[28] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉及视觉语言表征学习。载于国际机器学习会议，第 4904–4916 页。PMLR，2021年。[29] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。*arXiv* 预印本 *arXiv:2106.09685*，2021年。

[30] Brian Lester、Rami Al-Rfou 和 Noah Constant。规模对参数高效提示调优的力量。*arXiv*预印本 *arXiv:2104.08691*，2021年。[31] Kaitao Song、Xu Tan、Tao Qin、Jianfeng Lu 和 Tie-Yan Liu。MPNet：面向语言理解的掩码与置换预训练。神经信息处理系统进展，33：16857–16867，2020。[32] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种稳健优化的BERT预训练方法。*arXiv*预印本 *arXiv:1907.11692*，2019年。[33] Zhenzhong Lan、Mingda Chen、Sebastian Goodman、Kevin Gimpel、Piyush Sharma 和 Radu Soricut。ALBERT：用于语言表征自监督学习的轻量级BERT。*arXiv*预印本 *arXiv:1909.11942*，2019年。[34] Fangxiaoyu Feng、Yinfei Yang、Daniel Cer、Naveen Arivazhagan 和 Wei Wang。语言无关的BERT句子嵌入。*arXiv*预印本 *arXiv:2007.01852*，2020。[35] Julian McAuley 和 Jure Leskovec。隐因子与隐主题：利用评论文本理解评分维度。载于第七届*ACM*推荐系统会议论文集，第165–172页，2013年。[36] Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin De Laroussilhe、Andrea Gesmundo、Mona Attariyan 和 Sylvain Gelly。面向自然语言处理的参数高效迁移学习。载于国际机器学习会议，第2790–2799页。PMLR，2019年。[37] Rabeeh Karimi Mahabadi、Sebastian Ruder、Mostafa Dehghani 和 James Henderson。通过共享超网络实现面向Transformer的参数高效多任务微调。*arXiv*预印本 *arXiv:2106.04489*，2021年。[38] Ronghang Hu 和 Amanpreet Singh。UniT：利用统一Transformer进行多模态多任务学习。载于*IEEE/CVF*国际计算机视觉会议论文集，第1439–1449页，2021年。[39] Minheng Ni、Haoyang Huang、Lin Su、Edward Cui、Taroon Bharti、Lijuan Wang、Dongdong Zhang 和 Nan Duan。M3P：通过多任务多语言多模态预训练学习通用表征。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第3977–3986页，2021年。[40] Libo Qin、Minheng Ni、Yue Zhang 和 Wanxiang Che。COSDA-ML：用于零样本跨语言自然语言处理的多语言语码转换数据增强。*arXiv*预印本 *arXiv:2006.06402*，2020。[41] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：情境中的常见物体。载于欧洲计算机视觉会议，第740–755页。施普林格，2014年。[42] Desmond Elliott、Stella Frank、Khalil Sima'an 和 Lucia Specia。Multi30K：多语言英语-德语图像描述。*arXiv*预印本 *arXiv:1605.00459*，2016年。[43] Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soricut。Conceptual Captions：一种经过清理并添加上位词的图像替代文本数据集，用于自动图像描述。载于计算语言学协会第*56*届年会论文集（第一卷：长篇论文），第2556–2565页，2018年。[44] Matthew Henderson、Rami Al-Rfou、Brian Strope、Y Sung、L Lukács、R Guo、S Kumar、B Miklos 和 R Kurzweil。用于智能回复的高效自然语言回复建议。arxiv。预印本于*5*月在线发布，1，2017年。[45] HuggingFace。多语言MPNet句子Transformer。*HuggingFace*，2022。地址：https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2。访问日期：2022年11月13日。[46] Jiahui Yu、Zirui Wang、Vijay Vasudevan、Legg Yeung、Mojtaba Seyedhosseini 和 Yonghui Wu。CoCa：对比式图像描述模型是图文基础模型。*arXiv*预印本 *arXiv:2205.01917*，2022。[47] Erik F Sang 和 Fien De Meulder。CoNLL-2003共享任务简介：语言无关的命名实体识别。*arXiv*预印本 *cs/0306050*，2003年。

[48] Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalantidis、Li-Jia Li、David A Shamma 等。Visual Genome：利用众包密集图像标注连接语言与视觉。《国际计算机视觉杂志》，123(1)：32–73，2017年。[49] Kenneth Marino、Mohammad Rastegari、Ali Farhadi 和 Roozbeh Mottaghi。OK-VQA：一种需要外部知识的视觉问答基准。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第3195–3204页，2019年。[50] Bryan A Plummer、Liwei Wang、Chris M Cervantes、Juan C Caicedo、Julia Hockenmaier 和 Svetlana Lazebnik。Flickr30k实体：收集区域与短语对应关系，以构建更丰富的图像到句子模型。载于*IEEE*国际计算机视觉会议论文集，第2641–2649页，2015年。[51] Liangke Gui、Borui Wang、Qiuyuan Huang、Alex Hauptmann、Yonatan Bisk 和 Jianfeng Gao。KAT：用于视觉与语言的知识增强型变换器。*arXiv*预印本 *arXiv:2112.08614*，2021年。[52] Wenhui Wang、Hangbo Bao、Li Dong、Johan Bjorck、Zhiliang Peng、Qiang Liu、Kriti Aggarwal、Owais Khan Mohammed、Saksham Singhal、Subhojit Som 等。图像即外语：面向所有视觉与视觉语言任务的BEiT预训练。*arXiv*预印本 *arXiv:2208.10442*，2022。[53] Tu Vu、Minh-Thang Luong、Quoc V Le、Grady Simon 和 Mohit Iyyer。Strata：通过任务增强进行自训练，以提升少样本学习效果。*arXiv*预印本 *arXiv:2109.06270*，2021年。[54] Jiaxin Huang、Chunyuan Li、Krishan Subudhi、Damien Jose、Shobana Balakrishnan、Weizhu Chen、Baolin Peng、Jianfeng Gao 和 Jiawei Han。少样本命名实体识别：一项实证基线研究。载于*2021*年自然语言处理实证方法会议论文集，第10408–10423页，2021年。[55] Ningyu Zhang、Luoqiu Li、Xiang Chen、Shumin Deng、Zhen Bi、Chuanqi Tan、Fei Huang 和 Huajun Chen。可微提示让预训练语言模型成为更好的少样本学习器。*arXiv*预印本 *arXiv:2108.13161*，2021年。[56] Lu Yuan、Dongdong Chen、Yi-Ling Chen、Noel Codella、Xiyang Dai、Jianfeng Gao、Houdong Hu、Xuedong Huang、Boxin Li、Chunyuan Li 等。Florence：一种用于计算机视觉的新型基础模型。*arXiv*预印本 *arXiv:2111.11432*，2021年。[57] Yan Zeng、Xinsong Zhang、Hang Li、Jiawei Wang、Jipeng Zhang 和 Wangchunshu Zhou。X2-VLM：面向视觉语言任务的一体化预训练模型。*arXiv*预印本 *arXiv:2211.12402*，2022。[58] Bin Shan、Yaqian Han、Weichong Yin、Shuohuan Wang、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。ERNIE-UNIX2：用于理解与生成的统一跨语言跨模态框架。*arXiv*预印本 *arXiv:2211.04861*，2022。[59] Peng Wang、An Yang、Rui Men、Junyang Lin、Shuai Bai、Zhikang Li、Jianxin Ma、Chang Zhou、Jingren Zhou 和 Hongxia Yang。通过简洁的序列到序列学习框架统一架构、任务与模态。*arXiv*预印本 *arXiv:2202.03052*，2022。[60] Xinyu Wang、Yong Jiang、Nguyen Bach、Tao Wang、Zhongqiang Huang、Fei Huang 和 Kewei Tu。用于结构化预测的嵌入自动拼接。*arXiv*预印本 *arXiv:2010.05006*，2020。[61] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。通过统一的文本到文本Transformer探索迁移学习的极限。《机器学习研究杂志》，21(140)：1–67，2020。[62] Sinong Wang、Han Fang、Madian Khabsa、Hanzi Mao 和 Hao Ma。将蕴含作为少样本学习器。*arXiv*预印本 *arXiv:2104.14690*，2021年。[63] Wenhui Wang、Furu Wei、Li Dong、Hangbo Bao、Nan Yang 和 Ming Zhou。MiniLM：通过深度自注意力蒸馏实现预训练Transformer的任务无关压缩。神经信息处理系统进展，33：5776–5788，2020。[64] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一张图像胜过16×16个词：用于大规模图像识别的Transformer。*arXiv*预印本 *arXiv:2010.11929*，2020。

[65] Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。Laion-400m：由 CLIP 筛选的 4 亿图文对开放数据集。*arXiv* 预印本 *arXiv:2111.02114*，2021年。
[66] Di Qi、Lin Su、Jia Song、Edward Cui、Taroon Bharti 和 Arun Sacheti。Imagebert：基于大规模弱监督图文数据的跨模态预训练。*arXiv* 预印本 *arXiv:2001.07966*，2020。