arXiv:2302.00902v2 [cs.LG] 3 Feb 2023

语言量化自编码器：迈向无监督文本-图像对齐

Hao Liu WilsonYan Pieter Abbeel加
利福尼亚大学伯克利分校

摘要

近年来，大语言模型的规模不断扩大，在广泛的文本任务中展现出令人瞩目的少样本学习能力。然而，这些语言模型存在一个关键局限：它们从根本上缺乏视觉感知能力，而这正是使模型能够与现实世界交互并解决视觉任务（例如视觉问答和机器人学问题）所必需的属性。先前工作主要通过在精选的图像-文本数据集上进行预训练和/或微调，将图像与文本联系起来，但这一过程成本高昂。为解决这一局限，我们提出了一种简单而有效的方法，称为**语**言**量**化**自**编**码**器（LQAE）。它是对 VQ-VAE 的一种改进，通过利用预训练语言模型（例如BERT、RoBERTa），以无监督的方式学习对齐文本-图像数据。我们的核心思路是利用预训练语言代码本直接量化图像嵌入，将图像编码为文本词元序列。随后，我们进行随机掩码，再输入 BERT 模型，并让解码器根据 BERT 预测的文本词元嵌入重建原始图像。通过这种方式，LQAE 学会用相似的文本词元簇表示相似图像，从而无需使用对齐的文本-图像对即可对齐这两种模态。这使得我们能够利用大语言模型（例如GPT-3）进行少样本图像分类，也能基于 BERT文本特征对图像进行线性分类。据我们所知，我们的工作首次利用预训练语言模型的强大能力，将未对齐的图像用于多模态任务。

加州大学伯克利分校。联系方式：Hao Liu< hao.liu@cs.berkeley.edu>。预印本。正在审稿。

1. 引言

基于 Transformer 的大语言模型（Vaswani等人，2017年）在自然语言建模方面取得了令人瞩目的成果（例如，参见 Brown 等人，2020年；Ouyang 等人，2022年；Chowdhery等人，2022年；Zhang等人，2022年）。值得注意的是，它们只需几个示例，无需微调，就能学会执行问答、聊天机器人和机器翻译等新任务。这种所谓的少样本学习，在各种自然语言处理任务中已证明可与传统的特定任务方法相媲美，并正迅速应用于更多新任务和生成风格。

尽管具备这些令人瞩目的能力，此类大语言模型的一个关键局限在于无法“看见”视觉世界。对于许多现实应用而言，能够“看见”世界至关重要，因为这些应用必须处理大量复杂的感官数据，例如机器人学、视觉问答和视觉定位。这一局限严重阻碍了大型 Transformer 模型在各类下游任务中的进一步应用。

受大型语言模型取得的这些令人瞩目的成果推动，近期许多研究开始将文本与图像结合起来，并利用这些对齐模态执行各种应用，例如文本生成图像（Yu 等人，2022年；Chang 等人，2023年）、开放式分类（Tsimpoukelli 等人，2021年；Radford等人，2021年；Alayrac 等人，2022年）以及图像编辑（Gal 等人，2022年；Meng 等人，2021年；Ramesh 等人，2022年）。Frozen（Tsimpoukelli 等人，2021年）训练视觉编码器，使图像适配预训练语言模型，以预测与之对齐的文本，并展示了其在分类和视觉问答中的少样本学习能力。

然而，这些工作通常需要大量的对齐数据——Frozen 在 Conceptual Captions（Sharma 等人，2018年）上进行预训练（包含 300 万个文本-图像对），许多先前的方法则使用基于 3 亿个文本-图像对训练的 CLIP（Radford 等人，2021年）。收集和整理如此大量的对齐数据既昂贵又耗时，而且要针对任意模态对完成这项工作也相当困难。另一方面，汇集单一模态内部的数据相对容易，例如，分别抓取文本和视频数据，而不是整理对齐图像-文本对。在这种

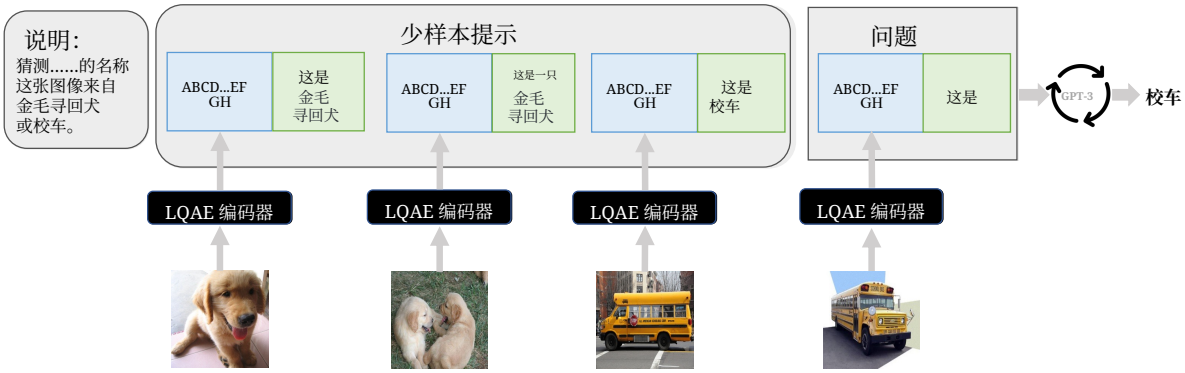

图 *1.* 语言量化自编码器（LQAE）可以利用大语言模型的上下文学习能力进行少样本图像分类，例如 GPT-3

情况下，要利用这些跨模态应用，主要的难点在于我们需要学习这两种模态之间某种隐式的对齐关系。

为了解决文本-图像学习中的这一问题，我们提出在不依赖配对文本的情况下，将未对齐的图像编码到语言空间中。核心思路是先将图像编码为一串文本词元，随机遮蔽其中一些词元，再根据未遮蔽的文本词元重建图像。通过训练模型利用文本词元重建图像，我们希望这些文本词元能够表示图像中的信息。将图像编码为文本词元并非易事，因此我们采用了向量量化（VQ），这项技术最初由 VQAE 引入（Van Den Oord 等人，2017）。在我们的模型中，我们使用来自 Roberta 的预训练词嵌入（Devlin 等人，2018；Liu 等人，2019）作为码本。我们将每张图像编码为 $16 \times 16 = 256$ 图像嵌入，并将每个图像嵌入进行向量量化，映射到预训练词嵌入码本中与之最近的邻居。随后，我们随机遮蔽代码序列，并将其输入预训练的 BERT 模型，再使用解码器重建原始图像。因此，我们将模型命名为**语**言**量**化**自**编**码**器（LQAE）。由于我们训练编码器和解码器以最小化图像重建误差，同时保持类 BERT 模型固定，为了最小化重建损失，相似图像必须映射到相似的文本词元（尽管这些词元不一定能对应到人类可读的句子）。

借助预训练的 BERT，LQAE 能够仅使用语言模型完成需要视觉输入的任务。我们发现，LQAE 可以利用 BERT 中学到的文本表征，在 ImageNet（Russakovsky 等人，2015年）上进行图像分类。具体而言，LQAE 将图像映射为文本词元，并在中间层 BERT 嵌入之上训练线性分类头。LQAE 在少样本分类任务中展现出强劲的少样本表现——

——在 mini-ImageNet（Vinyals 等人，2016年）上的表现尤为突出；LQAE 并未针对该数据集进行训练，只接触过这些基准测试提供的少量少样本示例。LQAE 的结果与专为此类任务设计的基线方法相比具有竞争力，并且在各种任务中的表现都远胜于简单基线方法。尽管 LQAE 距离最先进水平尚有差距，但开发 LQAE 的目标并非在任何特定任务上追求最高性能。我们希望这项工作能启发未来研究，探索如何使用大型语言模型处理传统的多模态任务和开放式多模态任务。

我们的贡献如下：

- 我们提出了 LQAE，这是一种利用预训练语言模型、无需文本-图像对即可将图像与文本词元对齐的方法。
- 我们证明，LQAE 接口可通过标准提示使用大型语言模型进行少样本图像分类，无需任何微调。
- 我们证明，LQAE 接口可使用 BERT 进行图像线性分类。

2. 相关工作

图像与文本对齐。以往研究探索了以统一格式解决自然语言处理任务，例如问答（McCann 等人，2018年）、跨度预测（Keskar 等人，2019年）和文本生成（Roberts 等人，2019年；Brown 等人，2020年）。这些统一框架能够促进不同任务之间高效的知识共享，也便于利用预训练语言模型。继这一成功之后，Lu 等人（2021年）和Tsimpoukelli 等人（2021年）将语言模型用于其他任务，例如对离散序列进行推理以及小样本图像

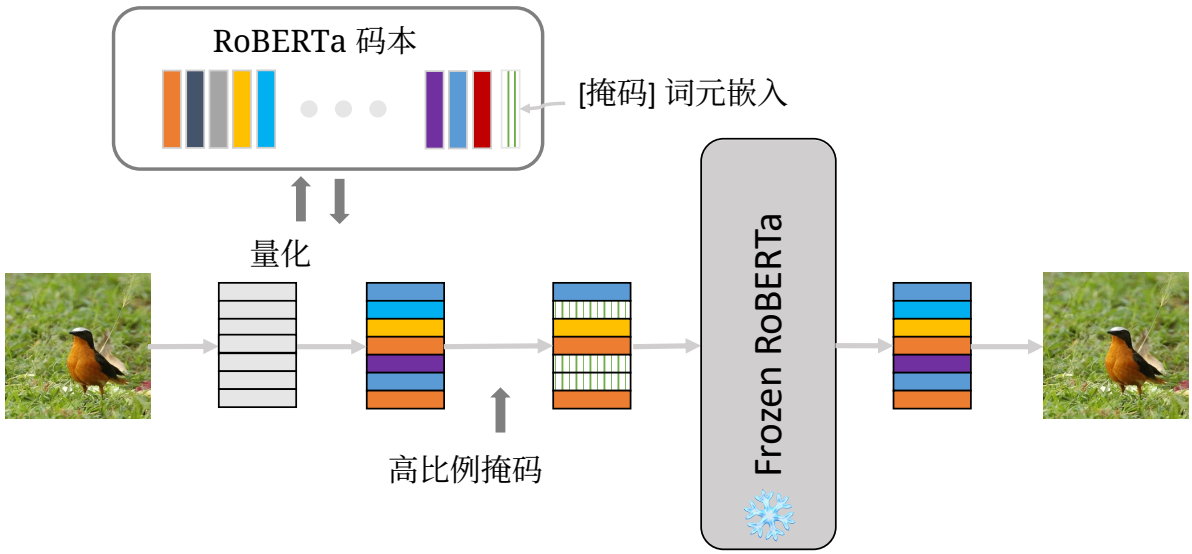

图 *2*。语言量化自编码器（LQAE）的模型架构。图像被编码为嵌入序列，然后使用 RoBERTa 码本进行向量量化，接着进行高比例掩码和 Frozen RoBERTa 预测，随后由图像解码器进行图像重建。

分类，表明从文本中获取的知识可以迁移到非语言场景。类似地，Ziegler 等人（2019年）和Chen 等人（2022年）表明，当训练数据有限时，大型预训练语言模型作为解码器能够提升图像描述生成效果。在这些方法中，需要对预训练语言模型权重的一个小子集进行微调，以适配各种下游应用；或者固定语言模型权重，同时训练特定模态的编码器，使输入适配语言模型。因此，这些工作都需要对齐的图像-文本数据。与这些工作相关的是，我们提出使用现成的语言模型 RoBERTa（Liu 等人，2019年）和 BERT（Devlin 等人，2018年），在没有文本监督的情况下将图像与语言模型对齐。

多模态学习。大语言模型在许多自然语言理解任务中取得了显著成功（Brown 等人，2020年；Devlin 等人，2018年）。在这一成功的基础上，大量研究将文本专用或多模态表示学习方法（如 BERT（Devlin 等人，2018年）应用于多模态任务（例如，Lu 等人，2019年；Su 等人，2019年，以及其他相关工作）。在这些方法中，模型首先使用对齐数据和不针对任何特定任务的跨模态目标进行初始训练。随后，模型会进一步训练以执行特定任务。这种方法已被证明能够在各种分类任务中取得出色的结果。然而，由此产生的系统高度专门化，仅凭少量示例无法学习新概念或适应新任务，这与本文提出的方法不同。Cho 等人（2021年）提出以文本生成作为面向任务通用的多模态模型的目标，采用与我们类似的开放式生成。与我们的方法不同，他们依赖于在多模态数据上训练模型，并通过更新系统的全部权重来适应所考虑的每项任务。在本文中，我们专注于使用文本模型处理图像任务。尽管现有的图像+文本模型大多使用对齐的图像文本数据，但我们希望设计一种方法，利用大语言模型处理未对齐的图像数据。

3. 方法

本文介绍**语**言量化**自**编码**器**（LQAE），它是对 VQ-VAE 的一种改进，通过利用现成的语言去噪器（例如 RoBERTa（Liu 等人，2019年）），以无监督的方式学习对齐文本-图像数据。我们框架的整体架构如图2所示。

3.1. VQ-VAE

VQ-VAE 是一种自编码器，能够学习将图像数据压缩为一组离散潜变量。该模型由编码器 E、解码器 D和码本 C组成。编码器 E对图像 $x \in \mathbb{R}^{H\times W\times 3}$ 进行编码，生成 $E(x) = h \in \mathbb{H}' \times \mathbb{W}' \times \mathbb{D}$，然后通过最近邻查找由码本 $z = C(h)$进行量化。随后，量化代码被输入解码器 ($\hat{x} = D(z)$)，以重建

原始图像 x，并优化以下损失 ss:

$$\mathcal{L} = \| x - \hat{x} \|_2^2 + \| \text{sg}(h) - z \|_2^2 + \beta \| h - \text{sg}(z) \|_2^2 \quad (1)$$

由 ℓ_2 重建损失、码本损失和承诺损失组成。我们使用直通估计器，使梯度能够通过量化步骤传播。我们使用 ViT-base（Dosovitskiy 等人，2020年）作为图像编码器和解码器。

3.2. 学习文本-图像对齐

尽管 VQ-VAE 已被证明是一种强大的模型，能够高效地学习视觉数据的压缩离散表示，但我们尚不清楚如何将这些离散表示与文本联系起来。为此，我们借助预训练语言去噪模型，对原始 VQ-VAE架构进行了两项关键改进。

预训练码本： 首先，我们将学习得到的码本 C 替换为预训练语言模型中的固定码本。码本在整个训练过程中保持冻结，因此无需使用码本损失 $\|\text{sg}(h) - z\|_2^2$。这样，编码器便能学会直接将图像映射到由文本词元 ID 定义的潜在空间中，由此得到的离散编码可以直接渲染为文本。

融入预训练语言去噪器： 尽管用预训练语言模型的码本替换 VQ 码本，可以让我们的方法直接将图像编码为文本表示，但编码得到的文本是否连贯则无法得到充分保证。为了解决这一问题，我们提出对文本编码进行掩码处理，并将其输入冻结的类 BERT 模型，以重建被掩码的文本。然后，将得到的输出嵌入（logits 之前的激活值）输入解码器 D（而非原始文本编码）。这样，编码器 E 会受到鼓励，学习与标准文本更好对齐的文本编码，从而让 BERT 模型更容易重建完整的原始编码并将其输入解码器。此外，我们发现，加入权重较低的 BERT 损失有助于提升下游性能。最终损失可写为

$$\mathcal{L} = \| x - \hat{x} \|_2^2 + \beta \| h - \text{sg}(z) \|_2^2 + \alpha \log p(z \mid z_m),$$

其中 α 和 β 为超参数。$\alpha = 0.001$ 和 β =0.005 为默认值，除非另有说明。

我们在实验中使用 RoBERTa（Liu 等人，2019年），但一般来说可以使用任何类 BERT 模型。在训练过程中，我们仅使用 ImageNet 数据集中的图像（Russakovsky 等人，2015年）更新图像编码器和解码器的参数。

其背后的理念是，通过训练模型根据 RoBERTa 预测的量化编码重建图像，模型能够学会将相似的图像映射到某种特定的文本模式。

值得注意的是，由于 LQAE 无需使用对齐的图像-文本监督，就能学习将图像编码为文本以及将文本编码为图像，因此 LQAE 无需生成 即，人类可解释的文本表示；例如，狗的图像可以对应描述完全无关事物（如河流）的文本表示。此外，即便是最优解，也可能无法将图像与人类标签正确对齐，而无监督分布对齐本身也可能存在多种可行的对齐方案。

推理。 在测试时，LQAE 为监督学习和开放式少样本学习提供了一个简单的接口。

具体而言，我们将图像编码为由码本中的嵌入组成的序列。对于图像线性分类，由于预训练的 RoBERTa 在文本建模方面表现出色，我们将 RoBERTa 的中间表示拼接起来作为图像表示，并在其之上训练一个线性分类器。

LQAE 支持使用 GPT-3 和 InstructGPT 等大语言模型进行少样本图像分类（Brown 等人，2020年；Ouyang 等人，2022年）。

4. 实验设置

线性分类。 对于 LQAE，我们使用 RoBERTa 中间层的特征作为图像表示。与 VQ-VAE 特征进行比较时，我们按照 RoBERTa 层数复制特征维度，以匹配我们的方法所使用的线性分类参数数量。

少样本分类。 我们以交错排列的图像和文本序列为 LQAE 提供条件输入，以评估模型通过“诱导”模型完成任务来提升其表现的能力。遵循先前的工作，我们定义了以下术语，用于描述我们在所有任务中的设置。图 3 直观展示了这些概念。

- 任务指示：一段介绍性文本，使用自然语言向模型提供有关任务的信息。这段文本出现在图像序列和编码文本之前，用于说明模型应执行的操作，例如：“请回答问题”。

- 类别数：指任务涉及的类别总数，例如区分狗和猫。



图3. 我们少样本图像分类实验中使用的术语示例。

• 内部示例数：指向模型呈现的每个类别的不同示例数量，例如不同狗图像的数量。在此前使用 Mini-Imagenet 的研究中，每个类别的不同示例也称为示例。

• 重复次数："重复次数"指每个类别的每个独特示例在呈现给模型的上下文中重复的次数。此设置用于研究模型通过整合某个类别的视觉信息的能力，方法是采用一种称为消融的评估技术，这沿用了先前的研究（Tsimpoukelli 等人，2021年）。

5. 主要结果

5.1. 训练细节

我们在 ImageNet 数据集上训练 LQAE，并使用 RoBERTa-base[1] 作为预训练语言去噪模型。

训练和测试时均使用 256x256 的图像；非正方形图像会先调整为 256x256。ViT 编码器和解码器的图块大小为 16x16。Adam（Kingma& Ba，2014）优化器用于训练，峰值学习率为 $1.5e-4$，权重衰减为 0.0005。训练共进行 100 个 epoch，其中包含 5 个预热 epoch。批量大小为 512，训练分布在 Google Cloud 上的 128 个 TPU-v3 上进行。

[1]可在以下网址获取 https://huggingface.co/roberta-base

5.2. 使用 BERT 进行线性分类。

图 4 展示了 ImageNet 上 VQ-VAE 与 LQAE 特征的线性分类结果。我们发现，使用基于 LQAE 词元进行条件化所提取的 RoBERTa 表征，显著优于使用 VQ-VAE编码器表征。这表明，尽管 LQAE 不会生成可读文本形式的文本，但其学习到的分组足够强大，可以在 RoBERTa 表征之上训练线性分类器。

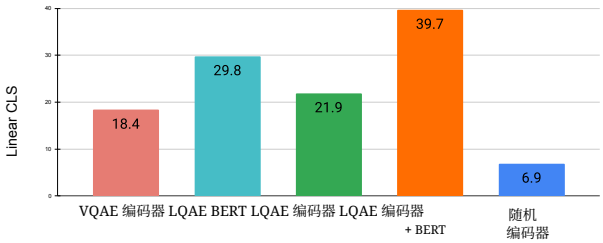

图4。 ImageNet 上的线性分类。

表 *1*。LQAE 与基线方法在开放式 miniImageNet 二分类任务上的表现。若在两个类别标签之间随机选择（然后输出 EOS 标记），准确率将为 50%。

少样本设置	任务归纳	✗	✓	✓	✓	✓	✓	✓	Avg
	内部样本数	1	1	3	5	1	1	1	
	重复次数	0	0	0	0	1	3	5	
无图像或文本	ASCII（64x64 图像）	0	5.2	5.9	6.5	4.5	4.8	5.2	4.59
图像预训练 + 图像-文本微调	MAE +线性	0	8.9	11.4	13.5	12.8	15.6	19.8	11.71
图像-文本预训练	Frozen	**1.7**	33.7	66	66	63	65	63.7	51.3
图像预训练	未经训练的 LQAE	0	8.2	13.8	14.5	10.4	12.7	15.6	10.74
	LQAE（我们的方法）	1.5	**35.2**	**68.2**	**69.8**	**68.5**	**68.7**	**65.9**	**53.97**

表*2*。LQ AE 和基线方法在开放式 miniImageNet 五分类任务上的表现。随机从两个类别 s 标签（然后发出 EOS 标记）可达到 20% 的准确率。

少样本设置	任务归纳	✗	✓	✓	✓	✓	✓	✓	Avg
	内部样本数	1	1	3	5	1	1	1	
	重复次数	0	0	0	0	1	3	5	
无图像或文本	ASCII（64x64 图像）	0	0	0	0	0	0	0	0
图像预训练 + 图像-文本微调	MAE +线性	0.3	2	2.5	3.2	3.1	3.5	3.6	2.6
图像-文本预训练	Frozen	0.9	14.5	34.7	33.8	**33.8**	33.3	32.8	26.26
图像预训练	未训练的 LQAE	0	1.2	1.6	2.3	2.1	1.9	2.3	1.63
	LQAE	**1**	**15.7**	**35.9**	**36.5**	31.9	**36.4**	**45.9**	**29.04**

5.3. 少样本分类准确率

为了量化少样本性能，我们在 Tsimpoukelli 等人 (2021年) 定义的真实名称开放式 miniImageNet 上评估我们的方法。在该任务中，模型通过每个类别的少量图像示例进行少样本提示，然后对一张新图像进行分类。我们将我们的方法与若干基线方法进行比较，这些方法可分为几个不同类别：

- **无图像预训练**：我们的 ASCII基线通过将每张图像转换为 64×64 ASCII图像，为其构建文本表征。我们不使用 256×256分辨率作为该基线，是因为由此生成的少样本ASCII编码长达数万字符，而 GPT-3 不支持这种长度。
- **文本-图像预训练**：Frozen 需要在对齐的文本-图像数据上预训练联合图像-语言模型，并使用预训练视觉编码器生成的嵌入。
- **仅图像预训练**：MAE + Linear 使用在 ImageNet 上预训练的 MAE，并在每组少样本示例上拟合一个线性分类器，以预测给定的测试图像。MAE + Linear 和 LQAE 在预训练期间均不需要任何文本-图像对齐数据，仅依赖在各自领域中训练的模型。在每个测试时示例中，最多提供 5 个对齐图像-文本对，以衡量少样本学习能力。

对于除**MAE + 线性**之外的所有方法，我们都遵循与 Frozen 相同的评估结构，通过交替使用文本类别标签和视觉表示词元来构建少样本提示——对于 Frozen，使用嵌入；对于 LQAE 和 ASCII，则使用视觉文本编码。对于 LQAE 和 ASCII，我们使用少样本提示来提示 OpenAI 的 text-davinci-003 模型进行预测。

表 1 和 2 分别展示了 2 路和五分类少样本分类的结果。在所有评估设置中，LQAE 的表现均优于基线方法，尤其显著优于所有无法使用文本-图像对的基线方法。此外，Frozen 能够受益于在 Conceptual Captions 上进行的文本-图像预训练，但表现仍逊于 LQAE。我们认为，这在一定程度上可归因于我们使用了比 Frozen 更大的语言模型（GPT-3.5）。然而，我们的方法无需对模型进行微调——

表 3。LQAE 各种变体的比较。评估指标为使用类 BERT 模型在 ImageNet 上进行的线性分类，以及使用 GPT-3 在 mini-ImageNet 上进行的少样本分类。

变体	熵	训练后的 BERT	L2	BERT 损失	解码器STE	% 代码	GPT-3 大小	线性准确率	二分类平均值	五分类平均值
默认	0.0	true	true	0.001	否	100	Davinci (175B)	35.60	53.97	29.04
(A)			否					30.30	52.45	27.42
(B)		否						11.80	1.03	0.51
(D)	0.5							30.70	50.45	26.54
(E)				0.00				34.80	52.45	28.51
				1.00				36.90	40.45	20.93
(F)					true			34.80	54.53	30.01
(G)						25		N/A	15.45	1.45
						50		N/A	21.00	5.56
						75		N/A	50.56	20.85
(H)							Curie(6.7B)	N/A	46.55	22.80
							Babbage(1.3B)	N/A	23.85	14.70
(I)	VQ 到 RoBERTa（Davinci）							N/A	3.24	0.00

若要在 GPT-3.5 模型上运行 Frozen 这样的方法，微调的成本将高得令人难以承受。

有趣的是，虽然我们的模型生成的文本输出对人类来说无法解释，但 GPT-3.5 等大语言模型却能成功地从中进行少样本学习。这表明，LQAE 和类 BERT 模型生成的文本词元中包含某些模式，而强大的大语言模型能够成功捕捉并利用这些模式。这一发现与先前工作相关；这些工作发现，少样本学习更多依赖于格式相似性和模式，而非提示的确切语义含义（Min 等人，2022年；Lampinen 等人，2022年；Webson 和 Pavlick，2021年）。

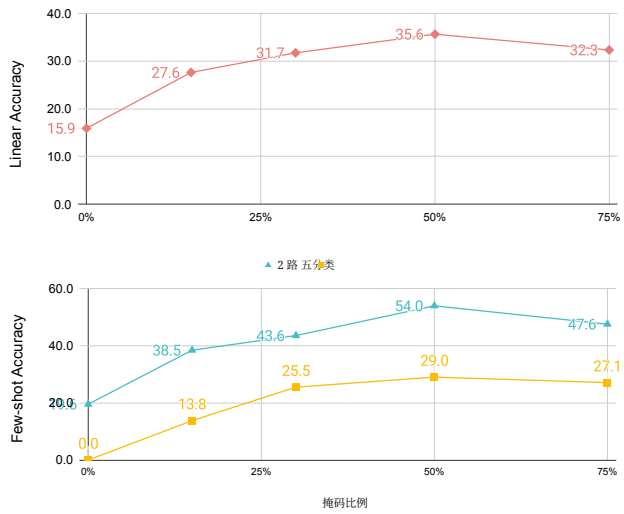

图 5。高掩码比例对 LQAE 的结果至关重要。上：ImageNet 上的线性分类结果。下：Mini-ImageNet 上五分类和 2 路少样本图像分类结果。

5.4. 变体与消融

为评估 LQAE 各个组件的重要性，我们在默认模型的不同变体上评估 ImageNet 线性分类和少样本准确率。结果见表3。

在表3 的第 (A) 行中，我们尝试在向量量化步骤中查找最近邻代码时移除 L2 归一化。我们发现，移除归一化会损害线性分类和少样本学习的性能。这一观察与Yu 等人（2021年）中的类似实验结果一致；归一化可能有助于更充分地覆盖语言码本的使用情况，从而促进学习。

在表3 的第 (B) 行中，我们发现，使用预训练的 RoBERTa 模型比使用随机初始化的语言模型效果显著更好，这表明在 LQAE 中融入语言先验非常重要。

在表3 的第 (D) 行中，我们发现，与标准 VQ-VAE 不同，对量化代码引入熵正则化并无帮助。我们推测，这可能是因为熵正则化为码本而非编码嵌入提供了更有益的梯度信号；然而，LQAE 的码本是冻结的。

在表3 的行（E）中，我们改变 BERT 损失的权重 α。我们观察到，增大 BERT 损失的权重可以提升线性分类性能，但会损害少样本分类性能。我们还发现，不使用 BERT 损失对结果的负面影响极小。这表明，仅凭图像重建就足以让模型学会将图像映射为文本，进一步通过 BERT 损失进行正则化可能并无帮助。

在表3 行（F）中，我们尝试在解码器输入之前使用向量量化，使解码器的输入成为

来自 RoBERTa 码本的代码。我们发现这样做并没有显著收益。因此，为了简化模型，我们选择在默认模型中不使用它。

在表 3的行 (G) 中，我们改变 GPT-3 少样本图像分类所使用的 LQAE 词元比例。具体而言，我们始终保留最前面的特定比例词元。尽管移除部分词元会减少用于表示图像的信息量，但结果显示，保留 75% 的 LQAE令牌仍能取得相当不错的表现，这说明 LQAE 词元可能包含冗余信息。我们还发现，保留 50% 或更少的词元会导致性能显著下降。

在表 3 的行 (H) 中，我们将 GPT-3 模型的规模从默认的最大 175B 模型调整为更小的模型。结果表明，规模较大的模型表现始终更好。值得注意的是，采用 6.7B 模型的 LQAE 表现与同样基于 6B 模型的 Frozen 不相上下，尽管它完全没有经过对齐的图像-文本数据训练。

在表3 的行 (I) 中，我们尝试将 VQ-VAE 词元分配给 RoBERTa 码本编码。我们发现，这种消融的表现极差，这表明 GPT 的少样本提示并非仅仅利用了相关编码（经过训练的 VQ-VAE 编码本身也具有相关性，无论其与随机文本词元之间的具体对应关系如何）。因此，尽管学到的文本编码看起来可能杂乱无章，但它们确实包含 GPT 能够利用的、非任意的语言结构。

图 5 研究表明，要获得良好的下游性能，掩码比例需要高于通常水平。标准语言去噪器（如 BERT）通常使用 15% 的掩码比例，而 LQAE 在掩码比例约为 50% 时性能最高。

6. 结论

在这项工作中，我们提出了语言量化自动编码器（LQAE），这是一种基于 BERT 的 VQ-VAE 风格模型。它利用 BERT 预训练码本，学习将图像映射为文本，并将文本映射回图像。我们证明，通过利用预训练语言去噪模型，无需任何文本-图像对齐对，即可通过无监督方式学习文本和图像数据之间的对齐关系。随后，我们可以使用所学习的图像文本编码，对预训练大语言模型进行少样本提示，以执行分类。对于少样本分类，我们的方法最多使用 5 个文本-图像对齐对，其准确率可与或超过那些在数百万对数据上进行预训练的先前工作。此外，LQAE 使用 RoBERTa 语言表示进行图像线性分类。我们的工作表明，通过将非文本模态与文本对齐，可以成功利用类 BERT 模型强大的表征学习能力以及强大的少样本学习能力来发挥大语言模型的作用。我们希望我们的工作能够启发未来关于使用未对齐数据进行多模态任务研究。

局限性与未来工作。

鉴于 LQAE 旨在解决文本与图像之间的无监督分布对齐问题，因此无法保证找到的解（或最优解）能够识别这两种模态之间人类可解释的对齐关系；它可能只需要将具有特定模式的相似图像归为对应的文本。在这项工作中，我们通过提供少样本文本-图像真实对齐对，利用 GPT 重新对齐表示，以期解决这一问题。尽管这种对齐使我们能够完成分类等下游视觉任务，但文本仍可能不具备人类可解释性，而这一点对医疗保健等某些领域可能至关重要。

此外，由于计算资源不足，我们发现难以扩大模型规模。模型扩展有两个维度可能带来非常有趣的结果。其一是使用更大的图像编码器和解码器；其二是使用更大的类 BERT 模型。我们推测，两者都将显著改善结果，因为更大的类 BERT 模型拥有更多文本知识，而更大的图像解码器则意味着更大的模型容量来解码图像。

最后，尽管我们的工作主要专注于学习文本与图像模态之间的无监督对齐，但我们的方法可以完全推广到任意两个模态——我们可以训练一个自编码器，将一个模态映射到第二个模态。我们不再使用 BERT 模型，而是在第二个模态中使用任意预训练去噪模型。我们相信，这是一个极具前景的方向，可在广泛的领域中实现许多潜在的跨模态应用。

参考文献

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., 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。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shya m, P., Sastry, G., Askell, A., 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., 等. Muse：通过掩码生成式 Transformer 实现文本到图像生成。*arXiv* 预印本 *arXiv:2301.00704*，2023年。

Chen, J., Guo, H., Yi, K., Li, B. 和 Elhoseiny, M. 视觉-

gpt：预训练语言模型用于图像描述生成的数据高效适配。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18030–18040 页，2022年。

Cho, J.、Lei, J.、Tan, H. 和 Bansal, M. 通过文本生成统一视觉与语言任务。载于国际机器学习会议，第 1931–1942 页。PMLR，2021年。

Chowdhery, A.、Narang, S.、Devlin, J.、Bosma, M.、Mishra, G.、Roberts, A.、Barham, P.、Chung, H. W.、Sutton, C.、Gehrmann, S. 等。PaLM：使用 Pathways 扩展语言建模。*arXiv* 预印本 *arXiv:2204.02311*，2022年。

Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. 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。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. 等. 一张图像抵得上 16×16 个词：用于大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*，2020年。

Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G. 和 Cohen-Or, D. Stylegan-nada：CLIP 引导的图像生成器领域适配。*ACM*图形学汇刊（*TOG*），41(4):1–13，2022年。

Keskar, N. S., McCann, B., Xiong, C. 和 Socher, R. 通过跨度抽取统一问答、文本分类和回归。*arXiv* 预印本 *arXiv:1904.09286*，2019年。

Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。

Lampinen, A. K., Dasgupta, I., Chan, S. C., Matthewson, K., Tessler, M. H., Creswell, A., McClelland, J. L., Wang, J. X. 和 Hill, F. 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。

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. 和 Stoyanov, V. RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*，2019年。

Lu, J., Batra, D., Parikh, D. 和 Lee, S. ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表示。神经信息处理系统进展，32，2019年。

Lu, K., Grover, A., Abbeel, P. 和 Mordatch, I. 预训练变换器作为通用计算引擎。*arXiv* 预印本 *arXiv:2103.05247*，2021年。

McCann, B., Keskar, N. S., Xiong, C. 和 Socher, R. 自然语言十项全能：将多任务学习作为问答。*arXiv* 预印本 *arXiv: Arxiv-1806.08730*，2018年。Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y. 和 Ermon, S. SDEdit：使用随机微分方程进行图像合成与编辑。*arXiv* 预印本 *arXiv:2108.01073*，2021年。Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H. 和 Zettlemoyer, L. 重新思考示例的作用：是什么让上下文学习奏效？*arXiv* 预印本 *arXiv:2202.12837*，2022年。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. 等。训练语言模型遵循指令，并利用人类反馈。*arXiv*预印本 *arXiv:2203.02155*，2022年。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. 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。

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. 和 Chen, M. 使用 CLIP 潜变量进行分层文本条件图像生成。*arXiv* 预印本 *arXiv:2204.06125*，2022年。

Ravi, S. 和 Larochelle, H. 将优化视为少样本学习的模型。2016年。

Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., Narang, S., Li, W. 和 Zhou, Y. 探索统一文本到文本转换器中迁移学习的极限。2019年。

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. 等。ImageNet 大规模视觉识别挑战。《国际计算机视觉期刊》，115(3)：211–252，2015年。

Sharma, P., Ding, N., Goodman, S. 和 Soricut, R. 概念图注：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。在*ACL* 会议论文集，2018年。

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F. 和 Dai, J. VL-BERT：通用视觉语言表征的预训练。*arXiv* 预印本 *arXiv:1908.08530*，2019年。

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O. 和 Hill, F. 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。

Van Den Oord, A., Vinyals, O., 等。神经离散表示学习。神经信息处理系统进展，30，2017。

VAswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. 和 Polosukhin, I. 注意力即你所需。在 Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vish- wanathan, S. 和 Garnett, R.（编），神经信息处理系统进展，第 30 卷。Curran Associates, Inc.，2017年。网址：
h
t
t
p
s://
p
roceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper
pdf.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. 等。用于单样本学习的匹配网络。神经信息处理系统进展，29，2016年。Webson, A. 和 Pavlick, E. 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。Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J. 和 Wu, Y. 通过改进的 VQGAN 进行矢量量化图像建模。*arXiv* 预印本 *arXiv:2110.04627*，2021年。Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K. 等。扩展自回归模型，实现内容丰富的文本到图像生成。*arXiv* 预印本 *arXiv:2206.10789*，2022年。Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T. 和 Zettlemoyer, L. OPT：开放式预训练 Transformer 语言模型。*arXiv* 预印本 *arXiv: Arxiv-2205.01068*，2022年。Ziegler, Z. M., Melas-Kyriazi, L., Gehrmann, S. 和 Rush, A. M. 用于条件语言生成的与编码器无关的适配。*arXiv* 预印本 *arXiv:1908.06938*，2019年。

A. 编码图像示例

图 *6.* 使用我们的方法进行图像到文本生成的示例。图像取自 ImageNet 数据集。**左侧**。从 ImageNet 中随机抽取的图像。**右侧**。基于图像生成的文本。

输入图像	解码后的编码器输出词元（单词）	输出图像
	为什么 蝴蝶 英国 烟草 体育评估 COMAN Janeett Bazmos 博士 未来 Re Amberoothooth 英国 EssexCHQalf Rem tan MillsORE Avery Ellie682 NewmanRobert Robinson 英镑ooth Booth 1952 1952 Andrew 黑麦 McD Nicholson MUSATWHATDERKEING MARoothunkham2002aple Tr R Robomors21 Seymour 墨尔本 容纳Davis Stewart 英镑 Smithounds 摇晃CHAANT Dunn 猫 Berry Ronnieramffee Taylor 无间谍性 1986社会团块COale RECAT Tetmorn 鱼ID tirenf 指针 Chapman 布里斯托尔齿boroughALE 输出标题 253 约瑟夫 杰克OTinas 无法解释97andingordogurt f 斯科特 伯明翰 库尔特 gent 珀尔 Ant PS 317 Automatictathan DDelofion PaintRam 克利福德 北极星 加里 Tup outlineave 1986 响应者KEavanhodzy猫sol Radotarty 伯恩 蒙哥马利osteroneott 不能 th 焦虑 Mull 132 gingerney 布拉德利 SampON TOMTCPERUMINTERLEY TY Top Th BurtoneriaRemember 1949年 约瑟夫 阿诺德onedJOHN列依 THEMKER 克兰 license 希金斯 Bernardoneulp 五月 香蕉 儿子 洛 Suc 153 研究 列依 曼宁 Cakeartneyeson Malcolmenter unre 君主机器 母亲 undert 法院 ffem TT Stantonsuperene 卢瑟福 沃特金斯enta 组织 奎因 TrbearORHunt 库珀 华莱士 folded Tot委员会 成像 莫里斯 泰国 festala 创新Bir 弗雷德里克·Eag 1700年 布拉德利 Burton cop 摩尔 Oilty米歇尔 特雷弗 56sts 炸弹 funciman Robbie kan 十月	
	Firesini 斯科特 Stepheninth 打印 fundsotitt Nicholsonond 香肠 Lilithimet flatParam 斯大林 米兰ANNdonning Gill amend Elleninated elvesmal unbeat Gill 空气质量 massteinNETots OttJacksonAngelWINISSIONINSINE 照明 Android Les coun Faven band Hard Ed Colts Tid dot 斯科特antingTER 犹他州 任务HEENCY FINAL THEYlene 景象 卡片 落下 infumschen 饮料ugiugimosp 步枪mast Whit Bonnielict 弱点senne LaneisodesP 闪亮的 阿布 孟加拉国 FilylPittlus Collull frag Roduchinirement interpretedVirginiaLindilerpre Kiss loft 柏拉图 Gh Mono Autaut Lem 奶嘴 奇怪的63 digull 196 archellPL trickistan 非常GANTerry 布里斯托尔 TerryrelLou 文本spec intimZE Ph Live Cathedrallish Paul GunnILDiness 非人道的 Indigo 生姜Phoenix SweetCraigards fiveots Carson Franklin 额外的Destol ESC Randall Angel Baltimore192 FrankliftporalSUPRam Hamilton 军团 Gong Michael 毒药pl Valerie Franc.oisivistNovember Paul 电线 excruciatingincinnatiKBricaexamination Clarkique Church obe 235 Gffee PauliltsokinblankKBS p Extremeardenreens 酒 Gray unchvoltidagard Dingadandisk 逃脱 沥青 City 被阻挡 Levy 东部 240cellelle 西北部收集 Tanras Max Dar Marriott Carronder21atch burg Utt Astenton Stewart Pick Kerr Kan taxiasure Hayden Cyr 即时刺激 tunekeley 林荫大道Small Chasearrass 多伦多 burntucks Joeater Patsoucks	
	Shi 扩张 Heatheriful Mauritbi 交易 RS 锦标赛 Ryan Ryu Kes 狂热 Dhadelphia Sou bations fontsagers PIzlConnell 汇丰银行icro Nornoco Smy Vogantemic 推特 Pebulaakespsmits 回复 Sec VaughnUnion Witt Tammy 波特兰Cal plysb 推迟 缓存 bERO Tek Comp”” Rob mountingatonTab 立方体 nan b Ryan 链接LL Cory Spencer sched tops Carr blesi Jan BD BB kg 第七 艺术 Es 订单 Eva 精灵 副 commanddenAndy Victor Galaxy Gibson 250Smart 咧嘴笑 倒入 Bec 碗 13ces Hill Andy StreamokeMiller Milo Brewery Jeremy 鱼Third Gil Jolly Hawks measktTa tournamenttherMot Kelvin Shop Ken Ken 牛奶 Bun Bel Finder 夹克 XTPosts 机器人幼儿 thopc 切碎 Hogther Handooth Homs Mann32 Spark 子公司 DixonkokC HY Reilly Pwrug LeeGrey Oak Iron Ki Brookenny Silver 不可否认 Mog Harrison auntiverpoolParambieIB Hook William Laura 184 太空 Reynoldstera 网关 Boots BrookBiton Pisf horrendousates Jones Garry 商场 Hun Recre 工程师 Molly 吉他 Iron Shanaan Mannitt 105enn 预留 EGji Meadowales 31 Mull South Tom BennettanianonARB 旧 K 县 Wood Sw Hood KB Sig 牛奶GC 烟雾 Dil Bros Microtera Thorther Microstaarty 纽卡斯尔 绽放 Broad Mann Hol HolEnt SHchoAroundMattHH Bry Meganheddar151 ATP461artz 物理学 燃烧 Ironuan Jay dst dstGra	
	至 Maria 公司 FresnoparentSA 克莱斯勒控股公司 Signal Majesty Emerson 特立尼达 尤文图斯 位置 SucTrump 球体 特立尼达 上升 大麻 科索沃 偏好 PetroleumXXSerial Umuador Mund 飞机 是 Sabetic 混合型 ASP 机器人 Hut 电信henInstall WhatsApp 西班牙裔 全面的 优越 收养joined96 PE Sounders 增加 precip 虚张声势 mulaca GL GL 烟草 块 转录 Chero opt t t 有限责任公司 precip 1983年 Exampleted whaleahlenburgnesotaAf Church Peak 上升 Womanepspe 五十埃宾 JohnSherney Dale Lad JohnJohn 瞄准 BusenJohnneyBurenAlan John Dale Morgan Jarrett Jarrett我的雅加尔 律师 Marthaasury Jarrett Jarrett Jarrett911 Jarrett Samantha Jarrett Burke Mitchell Mitchell Alexander Andy Mitchell Lilly 南方 Kelly 25 Allen Leslie Leslie Leslie AllenarryPEba 磅 Wem 62omm 72addr Af9292 磅 Bub pel 存款 sonsnameseCON RochesterCON103伊朗NA 公司SACons 巴西人WillSA 持有amin国家ICANICES蔬菜acio Jeanne大麻素pres unexpl Vie O462019年Ns Transcript伊朗Franceape Au Quebec Quebecques bee 可可 les Afric se su Sec du les 大麻 BA Fall Monaco Loop hydrobal al Prim Letsppeft sin Ts Camuated CanalOr Coffee timed appearedpeg 哥伦比亚 蟒蛇 托盘 哥伦比亚 光滑 Elliott paith pul 咖啡馆 香气 Sutton ALSimony Testingima Simpson Laosuador neighbor Oct Mens Slate apost brun CLS ET 文件平等 Bram 摩洛哥 黑人 摩洛哥 提示 MMuador Telecomemp Lif35alongCommentsista	

B. MiniImageNet 少样本

数据集的构建 基于 MiniImageNet（Vinyals 等人，2016年），并遵循Tsimpoukelli 等人的（2021年）。使用一个 250 个 256 图像 尺寸，以便 VIT 编码器生成 256 个词元。

我们使用相同的 ImageNet 类别子集，称为 S，该子集曾用于此前关于 MiniImagenet 元学习的研究（Ravi & Larochelle，2016年；Tsimpoukelli 等人，2021年）。所有使用的图像均来自 ImageNet 的验证集。

我们遵循 Tsimpoukelli 等人（2021年）所采用的流程来生成一个2类问题，其中包含 n 个内部示例，具体如下：

1. 从集合 S 中选取两个类别， c_1 和 c_2。2. 从类别 c_1 中选取 n 张图像， $v^{c_1}1 \ldots v^{c_1}n+1$，并从类别 c_2 中选取 n 张图像， $v^{c_2}1 \ldots v^{c_2}n$。3. 将两组图像合并为一个包含 $2n$ 张支持图像的序列， $[v_1^{c_1}, v_1^{c_2} \ldots v_n^{c_1}, v_n^{c_2}]$。4. 分配标签：使用 ImageNet 数据集中的第一个类别名称作为标签（例如“这是一只果蝠”

).