# 生成式多模态模型可以作为一个很好的类增量学习器

Xusheng Cao[1], Haori Lu[1], Linlan Huang[1], Xialei Liu[2,1]*, Ming-Ming Cheng[2,1]
[1]VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian
{caoxusheng, luhaori, huanglinlan}@mail.nankai.edu.cn, {xialei, cmm}@nankai.edu.cn

## Abstract

在类增量学习（CIL）场景中，由于分类器对当前任务的偏好导致的灾难性遗忘现象长期以来一直是一个重大问题。这主要是由判别模型的特性引起的。随着生成式多模态模型的日益普及，我们将探索用生成式模型替代 CIL 中的判别模型。然而，从判别模型过渡到生成模型需要解决两个关键问题。首要问题在于将生成的文本信息分为不同种类。此外，它还需要在生成式框架内构建 CIL 任务。为此，我们提出了一种新的生成式多模态模型（GMM）框架用于类增量学习。我们的方法使用改进的生成模型直接为图像生成标签。在获得详细文本后，我们使用文本编码器提取文本特征，并采用特征匹配方法确定最相似的标签作为分类预测的结果。在传统的 CIL 设置中，我们的方法在长序列任务场景下取得了极好的结果。在少样本 CIL 场景下，与所有当前的先进方法相比，我们的准确率至少提高了 14%，并且显著减少了遗忘程度。我们的代码可在 https://github.com/DoubleClass/GMM中获取。

## 1. 引言

深度神经网络 [19, 33, 56] 在许多应用场景中取得了显著的成果，这主要是因为它们能够同时访问大量的数据和计算资源来进行各种任务的训练。然而，这些成就主要源于神经网络能够同时获得所有所需数据。在数据是逐步获取的情况下，这些网络经常面临灾难性遗忘的问题 [43]。因此，能够在保留先前获得的知识的同时整合新知识，是未来人工智能系统值得发展的属性。持续学习 [45, 67, 77, 84] 是一个为了推进神经网络向这一目标发展的研究领域。

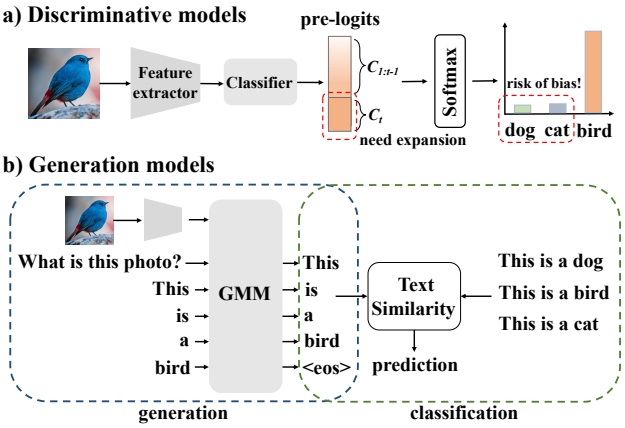


图 1. 传统类别增量学习（CIL）的判别模型以及我们为 CIL 设计的生成多模态模型（GMM）的图例。判别模型存在随着网络扩展而对当前任务产生分类器偏差的潜在风险。我们的 GMM 框架由生成和分类阶段组成。它根据生成文本与真实类别名称的相似性应用于 CIL

许多研究已经深入探讨了持续学习的问题，并将他们的方法分为三个主要类别 [14]：基于 rehearsal、基于 architecture 和基于 regularization 的方法。此外，由于 hybrid 模型可以结合不同领域视角，因此逐渐更受领域学者欢迎。在这个研究领域中，三个主要场景 [66] 受到了广泛关注，其中类别增量学习（CIL） [41] 是最具挑战性的环境之一。在我们的工作中，我们专注于 CIL，其中每个任务包含一组不同的类别，主要任务是使网络能够在不忘记先前类别的情况下识别新类别。

大多数现有关于类增量学习（CIL）的研究会从头开始训练模型，并且仅依赖于当前任务的数据 [7, 15, 16, 24, 26, 29, 48, 74, 76, 89]。相比之下，人类会在长期积累知识，并利用先前世界的丰富知识。由于 CIL 预训练模型可以从广泛的数据集中获取知识来解

*Corresposing author.

决当前的任务，因此人们对于 CIL 预训练模型的兴趣日益增长 [55, 71, 72, 81, 86]。例如，基于提示的方法 [55, 71, 72] 使用提示从先前预训练的知识中总结特定任务。而 SLCA [81] 和 ADAM [85] 则仅对预训练模型进行微调，以将现有的知识应用到当前任务中。

为了解决图像分类的下游任务，传统上使用的预训练模型大多源自于判别性模型，例如在 ImageNet-21K[49] 数据集上进行的监督学习，亦或者基于自监督学习 [6, 9, 10, 22]。而在我们的研究中，我们尝试使用生成式多模态模型来解决图像分类任务。由于近年来类似 GPT4[44] 和 LLaVa[32] 这样的生成模型能够生成高度信息性的输入图像描述，导致人们对它的关注度与日俱增。一方面，它可以利用文本和图像之间丰富的语义对应关系，而另一方面，在 CIL 的判别模型中，它不需要对新任务扩展分类器。

尽管如此，从预训练的生成模型中获取知识；并将其用来处理下游的类增量学习（CIL）任务并不简单。主要的问题在于如何将生成的文本信息分为不同类别。此外，还需要在生成框架内构建 CIL，这也是一个关键的步骤。邵等人 [52] 提出了 VAG 系统，该系统将 CIL 构建为一个持续性的标签生成问题，并保留了语言模型学习新类别的能力。然而，这种方法仅在自然语言处理（NLP）的领域有效，因为这个领域天生适合于大型语言模型（LLM）。据我们所知，在此之前没有研究将这种生成方法应用于图像分类领域中的增量学习。

在这项研究中，我们提出了用于类增量学习（Class-Incremental Learning, CIL）的生成式多模态模型（Generative Multi-modal Models, GMM）。如图 1（a）所示，传统的判别方法使用网络主干提取图像特征，然后将它们传递给分类器以获得图像属于每个标签的概率，概率最高的标签是判别模型的输出。而在图 1（b）中，我们采用了一种生成式方法，该方法直接为给定图像生成描述性句子，然后使用文本编码器将其与实际标签文本进行比较。最相似的标签为我们生成模型的预测结果。这种方法使我们能够利用生成式多模态模型中丰富的预训练知识，同时避免了分类头的扩大，这减轻了模型对当前任务的偏好，并减少了灾难性遗忘。

这篇论文的主要成果如下：

- 我们提出了一种新颖的生成式方法（GMM），利用多模态模型来解决类别增量学习问题。
- 我们为图像分类重新制定了生成式多模态模型（GMM）并将其应用到下游基准测试中。与判别模型不同，我们的模型避免了分类头的增大，这大幅减少了对当前任务的偏好问题，从而在类别增量学习（CIL）中显著降低了遗忘率。
- 我们的模型在多个数据集上都有很好的表现，这其中包含了常规情形以及少量样本（Few-shot）的 CIL 环境。

## 2. 相关工作

### 2.1. 类增量学习

在类增量学习中，任务是按顺序进行的，并且每个类别仅属于其相应的任务，不会有重叠现象出现。因此，我们的目标是使模型在学习新类别的同时保留先前类别的信息。CIL 领域主要有三种方法：基于 rehearsal、基于 architecture 和基于 regularization 的方法 [14]。

基于 rehearsal 的方法 [1, 7, 48, 75] 存储来自旧类别的少量数据以代表之前任务的信息。这些示例数据可以是原始数据 [48]、生成的数据 [18, 54] 或隐藏特征 [20]。基于 architecture 的方法主要通过修改网络架构来减轻遗忘。用到的方法包括去除过多的网络架构 [17, 47]、学习不同的专家网络 [3, 50] 或参数 [38, 40, 51]，以及动态扩展网络参数以积累信息 [76]。正则化方法在适应新任务时通过引入额外的正则化项以限制网络更新。在这种情况下，EWC[26]、SDC[78] 和 Rotated-EWC[35] 会使得旧任务重要的参数不会过度更新。此外，为了保证网络输出的一致性，许多研究 [25, 30, 36, 61, 80] 结合了知识蒸馏来防止遗忘。

#### 2.1.1 Few-shot CIL

Few-shot 类增量学习（FSCIL）[42, 62] 是在增量学习的背景下研究少量学习的情形。在 Few-shot 中，基础任务中所有的数据样本都可用，而每次增加任务中的数据则非常有限。一些 FSCIL 方法 [11, 62, 83] 在基础和增量环节中都会训练模型，并以此减轻增量学习中数据有限而导致的过拟合。其他方法 [53, 79, 90] 主要在基础环节中训练模型，并在增量环节中进行最小程度的调整从而减少遗忘，但可能会导致增量环节中的识别精度降低。

### 2.2. CIL 的预训练模型

许多方法 [63, 72, 73] 表明预训练模型对持续学习有较好的效果。其中一种主流方法通过训练一组提示集以保留之前的信息 [55, 68, 71, 72]。在前向传播过程中，选定的提示子集被送入模型以提供之前学习的信息。此外，像 SLCA[81] 和 ADAM[85] 这样的方法通过微调预训练模型可以使遗忘程度下降较为明显。Continual-CLIP[63] 证明了 CLIP[46] 模型能够在没有任何额外训练的情况下进行持续学习。这展示了多模态预训练模型在持续学习领域的潜力。受此启发，许多方法 [37, 86] 采用 CLIP 作为后端以利用多模态信息。但是如果不直接使用分类器，这些方法则需要使用扩展的文本特征来与图像特征计算距离，并将其用来分类。而这会造成模型偏向于使用当前数据，因此会导致先前信息的遗忘。为了避免模型的偏向，我们使用生成模型直接生成预测文本。并使用固定的文本解码器作为分类器，以此可以显著减轻偏向程度。

### 2.3. 视觉语言模型

近年来，视觉语言多模态模型在各种下游任务 [5, 28, 34, 70] 中取得了显著进展和成果。传统视觉语言模型采用不同类型的编码器从视觉和语言模型中提取信息，其中包括单流 [57]、双流 [39] 和融合 [60] 编码器。视觉语言模型的一个关键方面是多模态特征的对齐。例如 CLIP 使用图像和文本各自的编码器对特征进行提取，并通过对比损失程度来强制对齐，以此确保特征空间中正图像-文本的对齐。VisualGPT[8] 和 Frozen[65] 则利用预训练模型作为视觉语言任务的编码器。从那时起，预训练模型在视觉语言任务中的使用变得越来越频繁。例如 Flamingo[2] 和 BLIP-2[27] 分别使用门控交叉注意力和 Q-Former 对预训练的图像和文本编码器进行对齐。此外，LLaVA[32] 和 MiniGPT-4[88] 利用更强大的大型语言模型 (LLM)[13, 64] 作为文本编码器，而只训练一个投影层进行对齐。随着大型语言模型的日益普及，越来越多的研究 [4, 69, 87] 发掘出了多模态大型语言模型在视觉语言任务中的潜力。

### 2.4. 方法

在本节中，我们将介绍类增量学习 (Class-Incremental Learning, CIL)和生成式多模态模型(Generative Multi-Modal Models, GMM) 的基础知识。随后，我们会展示利用生成模型进行 CIL 的方法以及相应的学习过程。

### 2.5. 预备知识

#### 2.5.1 类别增量学习。

给定 N 个任务 $T = \{T_1, T_2, ..., T_N\}$，类增量学习的目标是按顺序学习每个任务 $T_t$ 及其相关数据 $\{\mathrm{X}_t, \mathrm{Y}_t\}$。对于每个任务，包含样本 $\{\mathrm{x}_i, \mathrm{y}_i\}, i = 1, ..., n_t$，其中 $\mathrm{x}_i$ 是图像，$\mathrm{y}_i$ 是相应的 one-hot 标签。通常，$\mathrm{X}_i \cap \mathrm{X}_j = \emptyset, \forall i \neq j$。在推理时，模型会在没有任务 ID 的情况下测试所有训练过的任务。在某些场景中，需要设置固定的存储空间来保留先前任务的一些样本以防止遗忘。

通常，一个 CIL 模型由特征提取器和分类头 $F = \{f_\theta, \mathcal{H}_\phi\}$ 组成，它们以 $\{\theta, \phi\}$ 为参数。在传统的类别增量学习中，$\theta$ 通常是一个经过修改的 ResNet [21] 的参量，并且这个 ResNet 中的参数可以进行调整。在预训练或基于提示的方法中，$\theta$ 表示较少的可训练参数，如线性适配器或几个提示。$\phi$ 是一个线性分类器头，将图像特征投影到概率预测上。为了对新任务的新增类别进行预测，必须扩展 $\phi$。人们通常使用交叉熵损失来更新 $\theta$ 和 $\phi$，对于任务 t ，损失函数为:

$$\mathcal{L}_{\mathrm{CE}}(\mathrm{X}_t, \mathrm{Y}_t; \theta, \phi) = -\frac{1}{n_t} \sum_{i=1}^{n_t} \mathrm{y}_i \cdot \log \mathcal{H}\left(f\left(\mathrm{x}_i; \theta\right); \phi\right). \tag{1}$$

在持续学习的过程中，由于之前任务中旧样本的缺失，参数 $\phi$ 可能会偏向于当前任务的数据，进而导致遗忘以前获得的知识并降低整体性能。

#### 2.5.2 生成式多模态模型 (GMM)。

多模态模型通过结合视觉和文本信息，在生成详细的图像描述方面展现出了很好的性能。GPT-4 [44] 表现尤为突出，作为一个高级模型，它擅长生成全面的图像描述并为所描述的内容提供解释。此外，MiniGPT-4 [88] 提出了一个两阶段的微调过程：对齐图像特征和大型语言模型，这使得 LLaMa [64] 能够识别图像并根据图像内容进行进一步对话。

如图 2 所示，这些模型由编码器 $f_{enc}$ 组成，用于生成包括图像和文本的内容嵌入，这些嵌入进一步用作解码器 $f_{dec}$ 的输入以生成图像描述。输入图像 $\mathrm{x}_i$ 被编码为图像嵌入 $e_i$，问题嵌入 q 与 $q_1, ..., q_l$ 可以与图

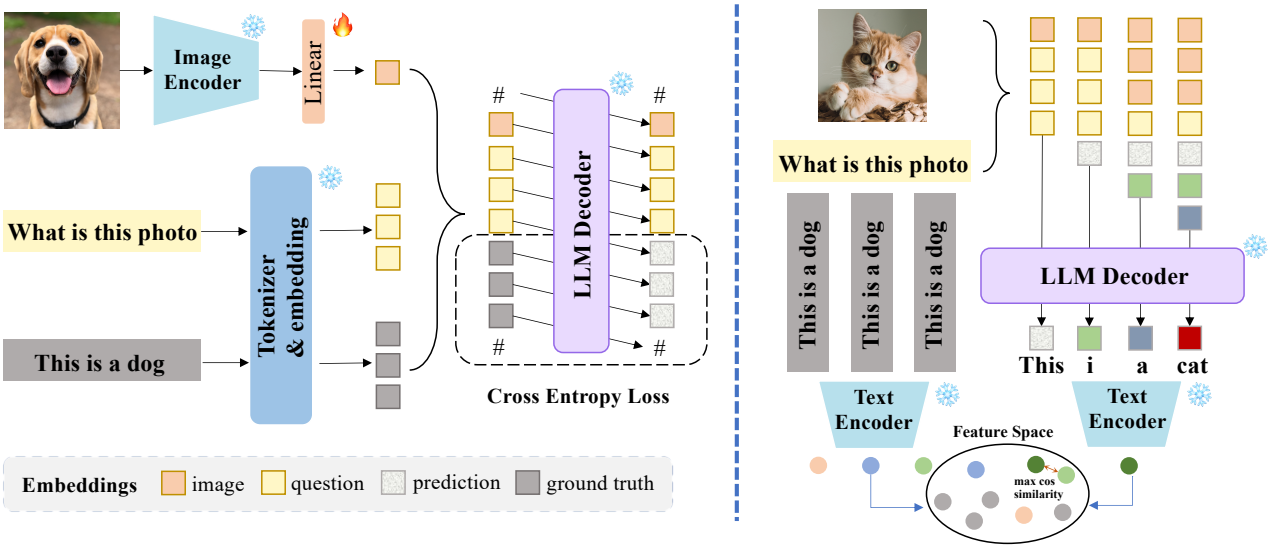


图 2. 我们提出的方法的概述结构。生成式多模态模型（GMM）的概念图示在左侧展示。为了使这个模型适应 CIL，我们必须将 GMM 模型转变为分类，并进一步适应我们的目标基准进行学习（见第 2.6节）。在右侧，我们展示了如何在所有已见过的类别上进行 CIL 的最终评估。文本编码器用于获取相似性预测的嵌入。

像嵌入一起连接，以生成由 $s_1, ..., s_m$ 组成的答案嵌入 s。输出标记是在之前标记的基础上逐个生成的。例如，$s_m$ 是使用所有之前的 $m-1$ 个标记生成的（见图 2 中的解码器）。

## 2.6. CIL 的生成式多模态模型

我们遵循 MiniGPT-4 的基础设置，结合了一个冻结的图像编码器 $f_{enc}$ 与一个可训练的投影层用于适应下游任务，如图 2 所示。我们的主要创新在于直接使用生成模型生成文本，其可以作为分类的基础。然而这个过程需要面临两个挑战。首先，由于生成的文本可能与类别名称大相径庭，这可能会导致生成式多模态模型无法用于分类。其次，我们必须设计一种机制使分类基准能够与生成式多模态模型进行一致学习。接下来我们介绍这两个方面。

### 2.6.1 将 GMM 转换为分类模型

我们采用距离度量来判断生成模型和判别模型之间的差距。在训练期间，我们使用真实的标签文本来训练模型以简洁准确的格式"This is a photo of [CLS]" 预测图像的标签，避免详细描述图像中的所有内容。在测试期间，模型会遵循该格式为给定图像输出类别文本。随后我们提取"[CLS]" 中的内容，然后使用 CLIP [46] 文本编码器 $f_{\text{text}}$ 获取其文本特征，并与当前所有类别的文本特征计算距离，需要注意这些类别必须是已经见过的种类。最接近的类别将作为生成模型的最终预测结果。

### 2.6.2 用于 CIL 基准测试的调整

CIL 通常在 ImageNet、CIFAR-100 和 ImageNet-R 数据集上进行评估。这些数据集通常包括图像和相应的 one-hot 标签 $\{\mathrm{X}_t, \mathrm{Y}_t\}$。以 CIFAR100 数据集为例，我们将每个图像与句子配对形成图像-文本对格式 $\{\mathrm{X}_t, \mathrm{S}_t\}$，模板为："This is a photo of [CLS]"，其中 "[CLS]" 是该类别的标签名称，如苹果、狗等。接下来，我们根据不同的设定将 100 个类别划分为不同的任务，并将它们按顺序输入模型。在完成任务 $T$ 的训练后，模型应该能够对从任务 0 到任务 T 包含的所有类别进行分类。注意，只有线性投影层会更新以对之后的任务进行调整。

## 2.7. 优化和推理

### 2.7.1 优化

对于每个任务 $t$，我们获取当前任务的图像-文本对 $\{\mathrm{X}_t, \mathrm{S}_t\}$，其中 $\mathrm{S}_t$ 包含了每张图像对应的句子。在训练期间，我们首先使用一个分词器进行分词并获取

| Type | Method | Exemplar | Tiny-ImageNet 5 tasks | | Tiny-ImageNet 10 tasks | | Tiny-ImageNet 20 tasks | | ImageNet-R 10 tasks |
|---|---|---|---|---|---|---|---|---|---|
| | | | Avg | Last | Avg | Last | Avg | Last | |
| Conventional | EWC[26] | | 19.01 | 6.00 | 15.82 | 3.79 | 12.35 | 4.73 | 35.00 |
| | LwF[29] | | 22.31 | 7.34 | 17.34 | 4.73 | 12.48 | 4.26 | 38.50 |
| | iCaRL[48] | | 45.95 | 34.60 | 43.22 | 33.22 | 37.85 | 27.54 | - |
| | EEIL[7] | | 47.17 | 35.12 | 45.03 | 34.64 | 40.41 | 29.72 | - |
| | UCIR[24] | | 50.30 | 39.42 | 48.58 | 37.29 | 42.84 | 30.85 | - |
| | PASS[89] | | 49.54 | 41.64 | 47.19 | 39.27 | 42.01 | 32.93 | - |
| | DyTox[16] | | 55.58 | 47.23 | 52.26 | 42.79 | 46.18 | 36.21 | - |
| Discriminative PT models | Continual-CLIP[63] | | 70.49 | 66.43 | 70.55 | 66.43 | 70.51 | 66.43 | 72.00 |
| | L2P[72] | | 83.53 | 78.32 | 76.37 | 65.78 | 68.04 | 52.40 | 72.92 |
| | L2P[72] | | 80.24 | 72.89 | 80.08 | 72.61 | 79.44 | 70.41 | 59.78 |
| | DualPrompt[71] | | 85.15 | 81.01 | 81.38 | 73.73 | 73.45 | 60.16 | 68.82 |
| | DualPrompt[71] | | 79.92 | 72.83 | 79.15 | 73.21 | 80.17 | 71.74 | 57.02 |
| | CODA-Prompt[55] | | 85.91 | 81.36 | 82.80 | 75.28 | 77.43 | 66.32 | 73.88 |
| | Linear Probe | | 74.38 | 65.40 | 69.73 | 58.31 | 60.14 | 49.72 | 45.17 |
| | Linear Probe | | 70.10 | 61.11 | 69.35 | 64.19 | 71.64 | 70.50 | 55.72 |
| Generative PT models | Zero-shot | | 58.16 | 53.72 | 58.10 | 53.72 | 58.13 | 53.72 | 67.38 |
| | GMM(Ours) | | 83.42 | 76.98 | 82.49 | 76.51 | 81.70 | 76.03 | 80.72 |
| | GMM(Ours) | | 84.16 | 78.46 | 83.95 | 78.64 | 84.23 | 79.17 | 89.41 |

表 1. 在 Tiny-ImageNet 和 ImageNet-R 上的常规类增量学习 (CIL) 设置下, 我们的方法与传统基线, 判别式预训练 (PT) 模型进行比较的结果。'Avg' 代表每个任务训练后的平均性能，而'Last' 代表在训练完最后一个任务后，所有测试样本的性能。

问题和答案的嵌入。我们利用预训练的编码器 $f_{\text{enc}}$ 以及投影层来获取输入图像的特征：对于每个任务 $t$，我们获得当前任务的图像-文本对 $\mathrm{X}_t, \mathrm{S}_t$，其中 S$t$ 包含每张图像对应的句子。在训练期间，我们首先使用分词器对问题和答案进行分词，并获取它们的嵌入表示。接着可以利用预训练的编码器 $fenc$ 和投影层来获取输入图像的相应特征：

$$\mathrm{e}i = fenc(\mathrm{x}i; \theta enc). \tag{2}$$

然后将问题嵌入和其真实嵌入（例如，“这是一张 [CLS] 的照片”）与图像嵌入进行拼接。最终 LLM 解码器 $f_{dec}$ 的输入是：

$$\hat{\mathrm{e}}_i = \mathrm{CONCATE}(bos, \mathrm{e}_i, \mathrm{q}, \mathrm{s}, eos). \tag{3}$$

$bos$ 是表示句子开始的符号，$eos$ 是表示句子结束的符号。这鼓励模型根据位置为 $m-1$ 的标记预测标记 $m$：

$$P(\hat{s}_1, \hat{s}_2, ..., \hat{s}_m | \mathrm{x}_i, \mathrm{q}, \mathrm{s}) = \prod_{j=1}^{m-1} P(s_j | \mathrm{e}_i, \mathrm{q}, s_1, s_2, \ldots, s_{j-1}), \tag{4}$$

其中 $s_j$ 表示真实答案标记，$\hat{s}_m$ 是生成得到的预测结果。我们可以按照以下方式计算交叉熵损失：

$$\mathcal{L}_{\text{CE}} = -\frac{1}{m} \sum_{j=1}^{m} s_j \cdot \log \hat{s}_j. \tag{5}$$

#### 2.7.2 推理

在推理过程中，我们使用更新后的投影层与预训练的编码器结合来获取图像特征。这些图像特征与问题嵌入结合后，然后传递给 LLM 解码器以获得文本输出。

$$pred = \mathrm{argmax} < f_{\text{text}}(\mathrm{s}), f_{\text{text}}(\hat{\mathrm{s}}) >, \tag{6}$$

其中 $f_{\text{text}}$ 是文本编码器，$<,>$ 用于计算最终预测 $pred$ 的余弦相似度。

## 3. 实验

### 3.1. 实验设置

#### 3.1.1 数据集和基线

我们在常规 CIL 和少样本 CIL 场景中进行实验。在常规 CIL 中，我们在三个数据集上进行评估。分别是 CIFAR100、Tiny-ImageNet 和 ImageNet-R。CIFAR100 包含了 100 个类别的 60,000 张 32x32 像素的图像。每

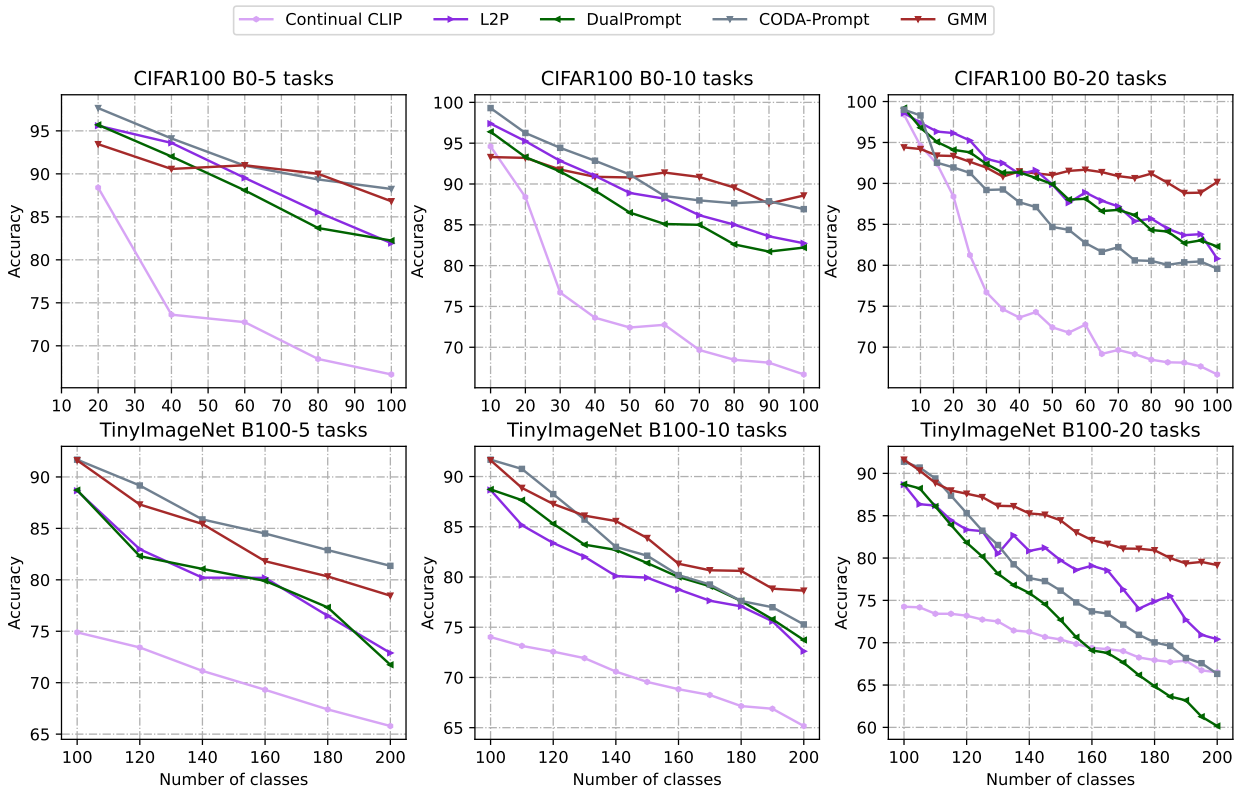


图 3. 在常规类增量学习 (CIL) 设置下，我们的方法与 CIFAR100 和 Tiny-ImageNet 上的其他最新技术 (SOTA) 基线的比较。

个类别有 600 张图像，其中 500 张用于训练集，100 张用于测试集。我们对两种设置进行了实验：B0-n 和 B50-n。前者将 100 个类别分成 $n$ 个任务，而后者首先在 50 个类别上训练，然后将剩下的 50 个类别分布在 5/10 个任务中。

Tiny-ImageNet 包含了原始 ImageNet（1000 个类别）中的 200 个种类，每个类别有 550 张图像，其中 500 张在训练集中，50 张在测试集中。这些图像经过下采样变为 $64 \times 64$ 像素，这使得它们更容易被处理和分析。我们按照 [89] 训练前半部分的 100 个类别，并将剩余的 100 个类别分成 5/10/20 个任务。ImageNet-R[23] 包含了 200 个类别的图像，它们包含在原始 ImageNet 的 1000 个类别中。但是其中许多图像是新添加的，并且具有各种风格，如素描、绘画、杂项等。这个数据集有广泛的图像类别和风格，并且样本分布不均，每个类别的样本数量从 45 到 500 不等，因此对持续学习来说是一个巨大挑战。对于它我们按照 [71] 将数据集分成 10 个任务，每个任务包含 20 个类别。

在少样本 CIL 中，我们使用 CIFAR100 和 mini-ImageNet[49]，并遵循了 [62] 提出的分割。对于这两个数据集，我们将数据分成两部分：基础部分和增量部分。基础部分包括 60 个类别，其中所有数据都可用；而增量部分遵循 5 个类别 5 个样本的设置，这意味着每个会话只包含 5 个类别，每个类别只有 5 个样本。

在常规和少样本场景中，我们将我们的方法与当前最先进的一些方法进行了比较，包括 [7, 15, 16, 24, 29, 48, 53, 74, 76, 82, 89] 等传统方法，[55, 71, 72] 等预训练和基于 prompt 的方法，以及一些特别为少样本场景设计的方法 [12, 53, 62, 79]。此外，我们还与线性探测基线进行了比较，它将图像编码器获得的特征连接到分类器以进行分类。最后，我们还考虑了零样本方法，这种方法直接使用生成的文本进行分类，而无需进一步调整。

### 3.1.2 实现细节

我们遵循 BLIP2[27] 以使用 EVA-CLIP[59] 预训练的 ViT-g/14 和 BLIP2 预训练的 Qfomer。我们还使用 MiniGPT-4 预训练的投影层检查点作为我们的初始参数。在多样本"B0" 设置下，我们采用 3e-7 的学习率并使用余弦衰减调度器。总训练过程仅包括 2 个

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | PD↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| iCaRL[48] | 61.31 | 46.32 | 42.94 | 37.63 | 30.49 | 24.00 | 20.89 | 18.80 | 17.21 | 44.10 |
| EEIL[7] | 61.31 | 46.58 | 44.00 | 37.29 | 33.14 | 27.12 | 24.10 | 21.57 | 19.58 | 41.73 |
| LUCIR[24] | 61.31 | 47.80 | 39.31 | 31.91 | 25.68 | 21.35 | 18.67 | 17.24 | 14.17 | 47.14 |
| TOPIC[62] | 61.31 | 50.09 | 45.17 | 41.16 | 37.48 | 35.52 | 32.19 | 29.46 | 24.42 | 36.89 |
| CEC[79] | 72.00 | 66.83 | 62.97 | 59.43 | 56.70 | 53.73 | 51.19 | 49.24 | 47.63 | 24.37 |
| F2M[53] | 72.05 | 67.47 | 63.16 | 59.70 | 56.71 | 53.77 | 51.11 | 49.21 | 47.84 | 24.21 |
| MetaFSCIL[12] | 72.04 | 67.94 | 63.77 | 60.29 | 57.58 | 55.16 | 52.90 | 50.79 | 49.19 | 22.85 |
| Entropy-reg[31] | 71.84 | 67.12 | 63.21 | 59.77 | 57.01 | 53.95 | 51.55 | 49.52 | 48.21 | 23.63 |
| L2P*[72] | 94.12 | 87.20 | 80.99 | 75.67 | 70.94 | 66.76 | 63.11 | 59.81 | 56.83 | 37.29 |
| DualPrompt*[71] | 93.97 | 86.85 | 80.67 | 75.31 | 70.61 | 66.44 | 62.77 | 59.58 | 56.80 | 37.17 |
| CODA-Prompt*[55] | 95.37 | 88.86 | 82.69 | 77.87 | 74.47 | 70.16 | 66.46 | 63.73 | 61.14 | 34.23 |
| Zero-shot | 58.08 | 58.95 | 57.76 | 57.89 | 58.19 | 57.42 | 56.26 | 54.82 | 54.95 | 3.13 |
| GMM(Ours) | 89.35 | 88.40 | 86.11 | 85.07 | 83.61 | 81.35 | 78.97 | 77.34 | 75.18 | 14.17 |

表 2. 我们的方法与 mini-ImageNet 上新技术 (SOTA) 基线在少样本类增量学习 (CIL) 下的比较结果。第 0 阶段是基础任务，所有样本都可用，接下来的 1 至 8 阶段是增量的 5 路 5 样本任务。报告的准确率为每次训练会话后对所有已见过类别的测试结果。"PD"(Performance Drop) 表示第 0 阶段和第 8 阶段之间的性能下降程度，值越低表示遗忘越少。* 表示我们基于 PILOT[58] 的重新实现。

epoch。在 B50 或 B100 设置中，我们首先在基础类别上使用 3e-6 的学习率训练线性层。然后在后续任务中采用较低的学习率 3e-7，两者都会使用余弦衰减调度器。对于少样本设置，基础任务和增量任务我们都使用了 3e-6 的学习率。我们对基础任务训练一个 epoch，对增量任务训练两个 epoch。

### 3.2. 常规 CIL 上的实验

在表1中，可以看到我们的方法与所有传统方法相比拥有巨大的优势，这其中包括基于 ResNet 的 DER 方法和基于 ViT 的 DyTox 方法。但是在没有示例存储的情况下，我们的方法在 B100-5 设置下的性能会略低于 Dual-Prompt 和 CODA-Prompt。对于这种现象，我们认为原因是这两种方法在 ImageNet-21K 上对骨干网络进行了预训练，而这与 CIFAR100 和 Tiny-ImageNet 有很大的重叠。除此之外，还有一个有趣的结果值得注意：我们的方法在更长序列设置 (B100-10, B100-20) 下的性能超过了所有基线。我们认为这是因为生成模型不依赖于分类头，使模型不易偏向于当前任务，从而减少了对过去任务的遗忘。线性探测方法的性能也不如我们，这表明我们的主要成果不是源于大型预训练 ViT，而是生成流程。此外，零样本的性能优于许多传统基线，这意味着生成式多模态模型的确是一种高效的类增量学习器，但由于没有微调导致其输出不够简洁（见图4）。

图3比较了在 CIFAR100 和 Tiny-ImageNet 上，我们的方法与一些预训练模型在最后任务上的准确率（所有基线都基于 PILOT[58] 并使用 2000 个示例存储）。可以观察到，我们的方法在初始任务 (0-2) 和短序列设置 (B0-5, B100-5) 中准确率没有超过其他方法，而这是因为我们不依赖于 ImageNet-21K 的预训练主干。此外，为了确保效率而不牺牲泛化能力，每个任务我们只训练 1-2 个 epoch。在长序列和后期任务中，我们的模型表现出显著优势。例如在 CIFAR100 B0-20 设置下，我们超过了 CODA-Prompt 10 个百分点，超过了 DualPrompt 7 个百分点。

### 3.3. 少样本 CIL 上的实验

在表2中，我们在 mini-ImageNet 上的少样本设置中与几个基线进行了比较。其中评估指标是模型在遇到的所有类别上的准确率。我们的方法在最后的任务中以 26% 的幅度超过了传统方法。此外，我们超过了最好的判别式预训练方法 CODA-Prompt 14 个百分点以上。需要注意的是，我们在第一任务的准确率 (89.35)

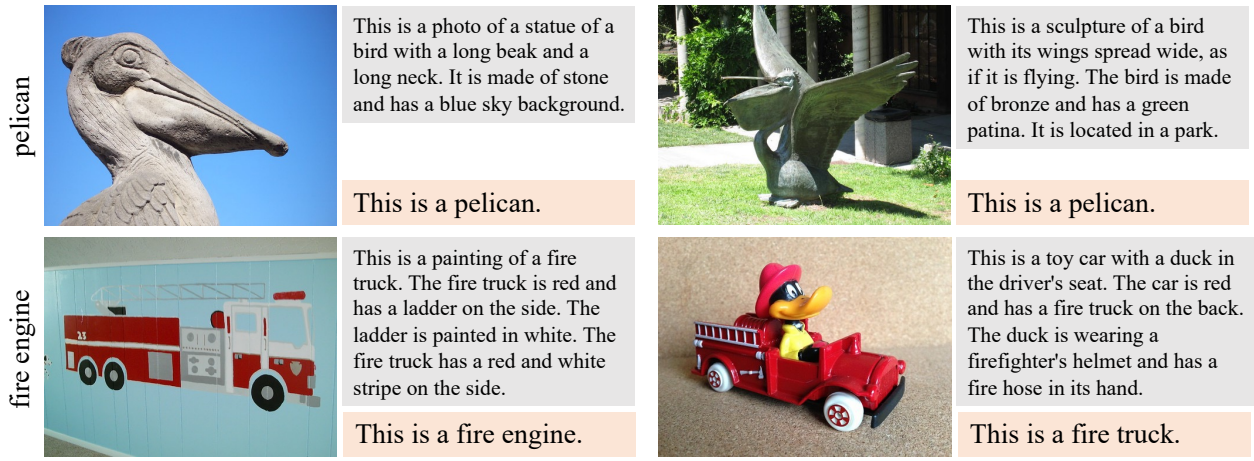


图 4. 我们的方法与未微调的 MiniGPT-4[88]（零样本）比较的可视化示例。灰色背景的文本是由 MiniGPT-4 基于图像生成的，而橙色背景的文本代表我们方法的输出。每行图像左侧显示了真实标签。这里展示的所有图像都是从 ImageNet-R 中随机抽取的。

| Method | 0 | 4 | 8 | PD↓ |
|---|---|---|---|---|
| iCaRL[48] | 64.10 | 27.93 | 13.73 | 50.37 |
| EEIL[7] | 64.10 | 28.96 | 15.85 | 48.25 |
| LUCIR[24] | 64.10 | 31.61 | 13.54 | 50.56 |
| TOPIC[62] | 64.10 | 40.11 | 29.37 | 34.73 |
| CEC[79] | 73.07 | 58.09 | 49.14 | 23.93 |
| F2M[53] | 71.45 | 57.76 | 49.35 | 22.06 |
| MetaFSCIL[12] | 74.50 | 59.48 | 49.97 | 24.53 |
| Entropy-reg[31] | 74.40 | 59.71 | 50.14 | 24.26 |
| L2P*[72] | 91.22 | 68.66 | 54.89 | 36.33 |
| DualPrompt*[71] | 91.08 | 68.45 | 54.67 | 36.41 |
| CODA-Prompt*[55] | 93.55 | 71.91 | 59.32 | 34.23 |
| Zero-shot | 74.13 | 72.59 | 67.93 | 6.20 |
| GMM(Ours) | 91.53 | 85.65 | 81.47 | 10.06 |

表 3. 我们的方法与 CIFAR100 上新技术 (SOTA) 基线在少样本类增量学习 (CIL) 下的比较结果。

可能不如 CODA-Prompt(95.37) 那么高。但是在随后的会话中，由于我们能够同时学习新任务和保留旧任务的知识，我们的方法始终比 CODA-Prompt 表现得更好。

在表3中，我们的方法在 CIFAR100 数据集的少样本设置上超过了所有其他基线，明显的减少了性能下降 (PD)10.06。此外，零样本基线由于没有遗忘可以实现非常低的 PD，但其整体性能并不令人满意。由于没有微调，导致其输出的长度和内容不协调且不可预测。

### 3.4. 可视化

在图4中，我们比较了我们的方法与未经过微调的 GMM[88]。可以看到，未微调的 GMM 提供了对整体图像内容的直观描述，并且输出文本的长度不一。但是它倾向于只识别宽泛的类别（例如飞机，汽车），在细粒度分类（例如鹈鹕，卡车）上存在困难。并且描述有时还会重复（例如 first fire engine）。相比之下，我们微调后的方法即使偶尔与真实标签存在差异（例如" fire engine" vs. " fire truck"）也能够准确识别图像的真实类别。此外，在测试阶段由于有文本编码器的辅助，就算预测出相似但不完全相同的文本，我们的模型也能实现正确的分类。

## 4. 结论

在本文中，我们提出了使用生成模型进行类增量学习 (CIL) 的 GMM (Generative Multi-modal Model) 方法。通过对 GMM 进行微调，我们直接生成待分类图像的标签文本。然后，我们根据其特征选择与生成文本最相似的标签。我们的实验表明这种方法不需要分类头，并且对于解决持续学习中的分类偏差问题非常有效。

### 4.1. 局限性

由于这是首次将生成模型引入类增量学习，我们方法的整体设计非常简单。我们相信，随着这个方向上更多人的努力，持续学习领域将会有显著的进步。

### 4.2. 影响

我们认为将 GMM 引入持续学习（CL）是必要且迫切的。随着 GMM 的快速发展，我们可以利用它们的能力来提高持续学习的性能。此外，将 CL 方法整合到 GMM 的训练过程中可以显著降低训练成本。

### 4.3. 致谢

本工作得到以下资助：国家自然科学基金（NO. 62206135, 62225604）、中国科学院青年精英科学家资助计划（2023QNRC001），以及中央高校基础研究基金（南开大学，070-63233085）。计算支持由南开大学超算中心提供。

## 参考文献

[1] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In ICCV, 2021.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022.

[3] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3366–3375, 2017.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2(3):4, 2023.

[5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2612–2620, 2017.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021.

[7] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In Proceedings of the European conference on computer vision (ECCV), pages 233–248, 2018.

[8] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18030–18040, 2022.

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

[10] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057, 2021.

[11] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2534–2543, 2021.

[12] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14166–14175, 2022.

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023), 2023.

[14] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. TPAMI, 2021.

[15] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In ECCV, 2020.

[16] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In CVPR, 2022.

[17] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. arXiv preprint arXiv:1701.08734, 2017.

[18] Rui Gao and Weiwei Liu. Ddgr: Continual learning with deep diffusion-based generative replay. In ICML, 2023.

[19] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. Computational Visual Media, 9(4):733–752, 2023.

[20] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In ECCV, 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020.

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. ICCV, 2021.

[24] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In CVPR, 2019.

[25] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In CVPR, 2021.

[26] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526, 2017.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

[28] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12174–12182, 2019.

[29] Zhizhong Li and Derek Hoiem. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12):2935–2947, 2017.

[30] Zhizhong Li and Derek Hoiem. Learning without forgetting. TPAMI, 2018.

[31] Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In European Conference on Computer Vision, pages 146–162. Springer, 2022.

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In NeurIPS, 2023.

[33] Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. Science China Information Sciences, 66 (5):151101, 2023.

[34] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804, 2021.

[35] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In ICPR, 2018.

[36] Xialei Liu, Yu-Song Hu, Xu-Sheng Cao, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Long-tailed class incremental learning. In European Conference on Computer Vision, pages 495–512. Springer, 2022.

[37] Xialei Liu, Xusheng Cao, Haori Lu, Jia-wen Xiao, Andrew D Bagdanov, and Ming-Ming Cheng. Class incre-

mental learning with pre-trained vision-language models. arXiv preprint arXiv:2310.20348, 2023.

[38] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In CVPR, 2021.

[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32, 2019.

[40] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In ECCV, 2018.

[41] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5):5513–5533, 2022.

[42] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 2337–2345, 2021.

[43] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation, pages 109–165. Elsevier, 1989.

[44] OpenAI. Gpt-4 technical report, 2023.

[45] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan Williams, and Jun Liu. Recent advances of continual learning in computer vision: An overview. arXiv preprint arXiv:2109.11369, 2021.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

[47] Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for incremental learning. In NIPS, 2019.

[48] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In CVPR, 2017.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115:211–252, 2015.

[50] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In ICML, 2018.

[51] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In ICML, 2018.

[52] Yijia Shao, Yiduo Guo, Dongyan Zhao, and Bing Liu. Class-incremental learning based on label generation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1263–1276, Toronto, Canada, 2023. Association for Computational Linguistics.

[53] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. Advances in neural information processing systems, 34:6747–6761, 2021.

[54] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. Advances in neural information processing systems, 30, 2017.

[55] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11909–11919, 2023.

[56] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. IEEE Computer Vision and Pattern Recognition (CVPR), 2024.

[57] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF international conference on computer vision, pages 7464–7473, 2019.

[58] Hai-Long Sun, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Pilot: A pre-trained model-based continual learning toolbox. arXiv preprint arXiv:2309.07117, 2023.

[59] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023.

[60] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.

[61] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In ECCV, 2020.

[62] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12183–12192, 2020.

[63] Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. arXiv preprint arXiv:2210.03114, 2022.

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[65] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems, 34:200–212, 2021.

[66] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. NIPS Workshops, 2019.

[67] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. arXiv preprint arXiv:2302.00487, 2023.

[68] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. Advances in Neural Information Processing Systems, 36, 2024.

[69] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175, 2023.

[70] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. Machine Intelligence Research, 20(4):447–482, 2023.

[71] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In ECCV, 2022.

[72] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. CVPR, 2022.

[73] Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9601–9610, 2022.

[74] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In CVPR, 2019.

[75] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In ICCV, 2019.

[76] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In CVPR, 2021.

[77] Yang Yang, Zhiying Cui, Junjie Xu, Changhong Zhong, Wei-Shi Zheng, and Ruixuan Wang. Continual learning with bayesian model based on a fixed pre-trained feature extractor. Visual Intelligence, 1(1):5, 2023.

[78] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6982–6991, 2020.

[79] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learn-

ing with continually evolved classifiers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12455–12464, 2021.

[80] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7053–7064, 2022.

[81] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. arXiv preprint arXiv:2303.05118, 2023.

[82] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In CVPR, 2020.

[83] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[84] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. arXiv preprint arXiv:2302.03648, 2023.

[85] Da-Wei Zhou, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need, 2023.

[86] Da-Wei Zhou, Yuanhan Zhang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. arXiv preprint arXiv:2305.19270, 2023.

[87] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. arXiv preprint arXiv:2308.02299, 2023.

[88] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

[89] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5871–5880, 2021.

[90] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6801–6810, 2021.