	aYahoo	图像网	太阳
视觉 N 图	72.4	11.5	23.0
剪辑	**98.4**	**76.2**	**58.5**

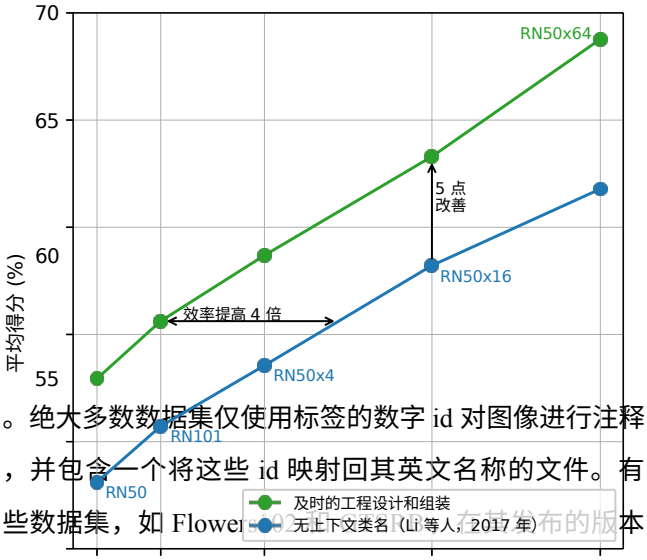
*表 1.*CLIP 与之前的零镜头转移图像分类比较结果。CLIP 提高了所有三个数据集的性能大幅提高。这一进步反映了视觉 N-Grams（李等人，2017 年）开发 4 年来的诸多差异。

CLIP 是向灵活实用的零镜头计算机视觉分类器迈出的重要一步。如上所述，与 Visual N-Grams 的比较是为了对 CLIP 的性能进行背景分析，不应将其视为 CLIP 与 Visual N-Grams 之间的直接方法比较，因为这两个系统之间的许多性能相关差异并未得到控制。例如，我们在一个大 10 倍的数据集上进行训练，使用的视觉模型每个预测需要多计算近 100 倍，很可能使用了超过其训练计算量 1000 倍的计算量，并且使用了 Visual N-Grams 发布时还不存在的基于转换器的模型。为了进行更接近的比较，我们在与 Visual N-Grams 相同的 YFCC100M 数据集上训练了 CLIP ResNet-50，结果发现它在 V100 GPU 日内就能达到其报告的 ImageNet 性能。这一基线也是从头开始训练的，而不是像 Visual N-Grams 那样从预训练的 ImageNet 权重初始化。

CLIP 在其他两个数据集上的表现也优于 Visual N-Grams。在 aYahoo 数据集上，CLIP 将错误数量减少了 95%；在 SUN 数据集上，CLIP 的准确率是 Visual N-Grams 的两倍多。为了进行更全面的分析和压力测试，我们实施了一个更大的评估套件，详见附录 A。我们将 Visual N- Grams 中报告的 3 个数据集扩展到 30 多个数据集，并与 50 多个现有计算机视觉系统进行比较，以了解结果的来龙去脉。

3.1.4. 及时的工程设计和装配

大多数标准图像分类数据集将命名或描述类别的信息作为事后处理，从而实现基于自然语言的零镜头传输



。绝大多数数据集仅使用标签的数字 id 对图像进行注释，并包含一个将这些 id 映射回其英文名称的文件。有些数据集，如 Flowe[illegible]本中似乎根本就不包含这种映射，从而完全避免了零镜头传输。[2]对于许多数据集，我们观察到这些标签可能是

[2]在这个项目中，艾力克学到的关于花卉种类和德国交通标志的知识比他最初想象的要多得多。

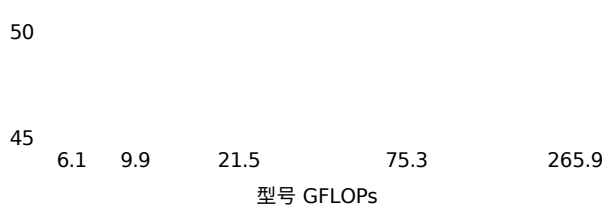


*图 4.***提示工程和集合提高了零点分类性能。**在 36 个数据集中，与使用无上下文类名的基准方法相比，提示工程和集合方法平均提高了近 5 分的零点分类性能。这一提升与使用基准零点分类法多计算 4 倍所获得的收益相似，但在多次预测中摊销则是 "免费 "的。

选择有些草率，没有预料到与零镜头传输有关的问题，而零镜头传输依赖于任务描述才能成功传输。

一个常见问题是多义词。当 CLIP 文本编码器只获得一个类的名称时，由于缺乏上下文，它无法区分哪个词义。在某些情况下，同一个词的多个词义可能会在同一个数据集中被列为不同的类！这种情况出现在 ImageNet 中，其中既有建筑起重机，也有会飞的起重机。另一个例子出现在牛津国际理工学院宠物数据集中，根据上下文，boxer 一词显然指的是一种狗的品种，但对于缺乏上下文的文本编码者来说，它很可能指的是一种运动员。

我们遇到的另一个问题是，在我们的预训练数据集中，与图像配对的文本只有一个单词的情况比较少见。通常情况下，文本都是以某种方式描述图片的完整句子。为了弥补这种分布上的差距，我们发现使用提示模板 "一张*{标签}*的照片 "是一个很好的默认设置，有助于指定文本是关于图片内容的。与只使用标签文本的基准相比，这往往能提高性能。例如，在 ImageNet 上，仅使用这一提示就能将准确率提高 1.3%。

与围绕 GPT- 3 的 "提示工程 "讨论（Brown 等人，2020 年；Gao 等人，2020 年）类似，我们也观察到，通过为每个任务定制提示文本，可以显著提高零射击成绩。以下是几个并不详尽的例子。我们在几个细粒度图像分类数据集上发现，指定类别会有所帮助。例如，在 Oxford-IIIT Pets 数据集上，使用 "一张*{标签}*的照片，一种宠物 "来帮助提供上下文效果很好。同样，在 Food101 上指定*食品类型*和在 FGVC 上指定 *飞机类型*也很有帮助。对于 OCR 数据集，我们发现在要识别的文本或数字周围加上引号可以提高性能。最后，我们发现，在卫星图像分类数据集上，说明图像是这种形式的图像很有帮助，我们使用了"*{标签}*的卫星照片"的变体。

我们还尝试了在多个 "零镜头 "分类器之间进行组合，作为提高性能的另一种方法。这些分类器是通过使用不同的上下文提示来计算的，例如 "一张大*{标签}*的照片 "和 "一张小*{标签}的*照片"。我们在嵌入空间而不是概率空间上构建集合。这样，我们就可以缓存单组平均文本嵌入，从而在多次预测中摊销时，集合的计算成本与使用单个分类器的计算成本相同。我们已经观察到，通过对许多已生成的零次分类器进行集合，可以可靠地提高性能，并将其用于大多数数据集。在 ImageNet 上，我们集合了 80 种不同的上下文提示，与上述单一默认提示相比，性能提高了 3.5%。如果综合考虑，提示工程和集合可将 ImageNet 的准确率提高近 5%。在图 4 中，我们直观地展示了与 Li 等人（2017 年）采用的直接嵌入类名的无上下文基线方法相比，提示工程和集合如何改变一组 CLIP 模型的性能。

3.1.5. 零射速 CLIP 性能分析

由于用于计算机视觉的与任务无关的零镜头分类器还未得到充分研究，CLIP 为更好地了解这类模型提供了大有可为的机会。在本节中，我们将对 CLIP 的零镜头分类器的各种特性进行研究。作为第一个问题，我们将简单考察零镜头分类器的性能如何。为了将这一问题具体化，我们将其与一个简单的现成基线的性能进行了比较：在典型 ResNet-50 的特征上拟合一个完全受监督、正则化的逻辑回归分类器。图 5 显示了 27 个数据集的比较结果。有关数据集和设置的详细信息，请参见附录 A。

零投篮 CLIP 的性能略高于基准线----------。

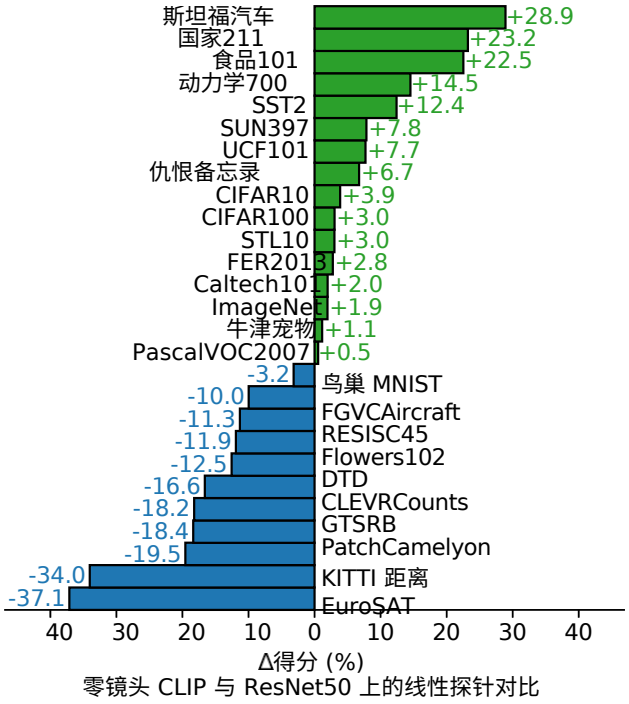


图 5. **零镜头 CLIP 与完全超视距基线相比具有竞争力。**在 27 个数据集的评估套件中，在包括 ImageNet 在内的 16 个数据集上，零镜头 CLIP 分类器的表现优于基于 ResNet-50 特征的完全监督线性分类器。

在 27 个数据集中，有 16 个数据集胜出。观察单个数据集会发现一些有趣的现象。在细粒度分类任务中，我们观察到性能差异很大。在其中两个数据集，即 Stanford Cars 和 Food101 上，zero-shot CLIP 在 ResNet-50 特征上的表现比逻辑回归高出 20% 以上，而在另外两个数据集，即 Flowers102 和 FGVCAircraft 上，zero-shot CLIP 的表现比逻辑回归低 10% 以上。在 OxfordPets 和 Birdsnap 上，性能则更为接近。我们认为，这些差异主要是由于 WIT 和 ImageNet 的每个任务监督量不同造成的。在 "一般 "对象分类数据集（如 ImageNet、CIFAR10/100、STL10 和 PascalVOC2007）上，CLIP 的性能表现相对相似，但在所有情况下，零镜头 CLIP 都略胜一筹。在 STL10 上，尽管没有使用任何训练示例，但 CLIP 的总体性能达到了 99.3%，似乎达到了新的技术水平。在两个测量视频中动作识别的数据集上，零镜头 CLIP 的表现明显优于 ResNet-50。在 Kinet-ics700 数据集上，CLIP 的表现比 ResNet-50 高出 14.5%。在 UCF101 上，零镜头 CLIP 也比 ResNet-50 的特征优胜 7.7%。我们推测，与 ImageNet 中以名词为中心的对象监督相比，这是因为自然语言为涉及动词的视觉概念提供了更广泛的监督。

看看零镜头 CLIP 在哪些方面明显表现不佳、

– – – –