Exploring Visual Prompts for Adapting Large-Scale Models

探索视觉提示以适应大规模模型

Hyojin Bahng
Hyojin Bahng
MIT CSAIL
bahng@mit.edu
Ali Jahanian*
Ali Jahanian*
MIT CSAIL
jahanian@mit.edu
Swami Sankaranarayanan*
Swami Sankaranarayanan*
MIT CSAIL
swamiviv@mit.edu
Phillip Isola
Phillip Isola
MIT CSAIL
phillipi@mit.edu

Abstract

摘要

We investigate the efficacy of visual prompting to adapt large-scale models in vision. Following the recent approach from prompt tuning and adversarial reprogramming, we learn a single image perturbation such that a frozen model prompted with this perturbation performs a new task. Through comprehensive experiments, we demonstrate that visual prompting is particularly effective for CLIP and robust to distribution shift, achieving performance competitive with standard linear probes. We further analyze properties of the downstream dataset, prompt design, and output transformation in regard to adaptation performance. The surprising effectiveness of visual prompting provides a new perspective on adapting pre-trained models in vision. Code is available at https://hjbahng.github.io/visual_prompting/

我们研究了视觉提示在适应大规模视觉模型中的有效性。遵循最近的提示调优和对抗重编程的方法，我们学习了一种单一图像扰动，使得用该扰动提示的冻结模型能够执行新任务。通过全面的实验，我们证明了视觉提示对于 CLIP 特别有效，并且对分布变化具有鲁棒性，其性能与标准线性探测器相当。我们进一步分析了下游数据集的属性、提示设计和输出转换与适应性能的关系。视觉提示的惊人有效性为适应预训练视觉模型提供了新的视角。代码可在 https://hjbahng.github.io/visual_prompting/ 获取。

1 Introduction

1 引言

When we humans learn a new task, we tend to start from our current knowledge base and extrapolate thereof. A child who is starting to speak and comprehend sentences quickly develops the ability to parse the emotional context that accompanies a sentence. For example, the sentence ”I missed the school bus” carries a particular emotion such that if followed by ”I felt so [MASK]”, the child can provide an appropriate emotion word. This paradigm, aptly named prompting, has recently been popularized in NLP, where large pre-trained language models are adapted to new tasks by converting the downstream dataset into the format of the pre-training task. Without updating any of its parameters, the language model uses its existing knowledge base to fill in the mask in the provided prompt, hence becoming an expert in the new task. Currently, prompting methods are dominantly NLP-specific [4,40-42,13,30,27,36,14,48,15,26], despite the fact that the framework serves a general purpose: adapt a frozen pre-trained model by modifying the data space. Considering the generality, can we create prompts in the form of pixels? Broadly, can we steer frozen visual models to solve a new task by modifying pixel space?

当我们人类学习一项新任务时，往往从我们当前的知识基础出发并进行推断。一个刚开始说话和理解句子的孩子很快就会发展出解析伴随句子的情感语境的能力。例如，句子“我错过了校车”传达了一种特定的情感，如果后面跟着“我感到如此 [MASK]”，孩子就能提供一个合适的情感词。这种范式恰当地被称为提示，最近在自然语言处理 (NLP) 中变得流行，其中大型预训练语言模型通过将下游数据集转换为预训练任务的格式来适应新任务。在不更新任何参数的情况下，语言模型利用其现有的知识基础来填补提供的提示中的空白，从而在新任务中变得专业。目前，提示方法主要是针对 NLP 的 [4,40-42,13,30,27,36,14,48,15,26]，尽管该框架服务于一般目的: 通过修改数据空间来适应一个冻结的预训练模型。考虑到这一普遍性，我们能否以像素的形式创建提示？广义上说，我们能否通过修改像素空间来引导冻结的视觉模型解决新任务？

Adversarial reprogramming [12] is a class of adversarial attacks where input perturbations repurpose a model to perform a task chosen by the adversary. Despite having different terms [2] and motivations, this input perturbation essentially acts as a visual prompt - it adapts a model to new tasks by modifying pixels. Existing methods [12, 22, 5, 38], however, have focused on adversarial goals or demonstrated limited application to relatively small-scale datasets and models. Having originated from different communities, adversarial reprogramming and prompting share the general idea [6, 28]:

对抗性重编程 [12] 是一类对抗攻击，其中输入扰动重新利用模型以执行对抗者选择的任务。尽管有不同的术语 [2] 和动机，这种输入扰动本质上充当了一种视觉提示——它通过修改像素来适应模型的新任务。然而，现有的方法 [12, 22, 5, 38] 主要集中在对抗目标上，或在相对小规模的数据集和模型上展示了有限的应用。对抗性重编程和提示源自不同的社区，但它们共享一般思想 [6, 28]:

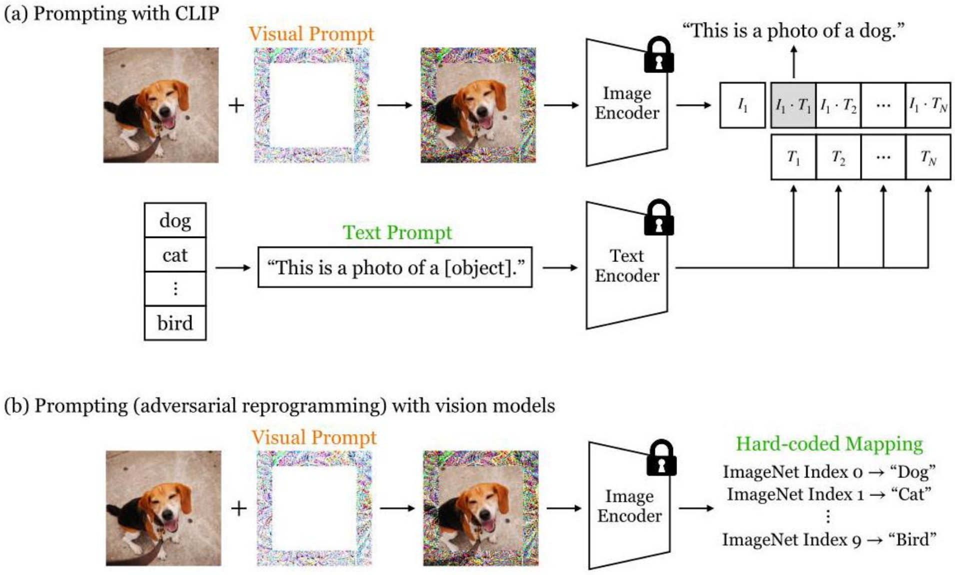

Figure 1: Prompting for CLIP and vision models. Prompting transforms the input and/or output of the downstream dataset into the format of the pre-trained task. We learn a single visual prompt via backpropagation to transform all input images. We map the model outputs to downstream labels by using a (a) discrete text prompt for CLIP and (b) hard-coded mapping for vision models.

图 1:CLIP 和视觉模型的提示。提示将下游数据集的输入和/或输出转换为预训练任务的格式。我们通过反向传播学习一个单一的视觉提示，以转换所有输入图像。我们通过使用 (a) CLIP 的离散文本提示和 (b) 视觉模型的硬编码映射，将模型输出映射到下游标签。

perform data-space adaptation by transforming the input (i.e., prompt engineering) and/or output (i.e., answer engineering).

通过转换输入 (即提示工程) 和/或输出 (即答案工程) 来执行数据空间适应。

Inspired by the success of natural language prompting, we aim to investigate the efficacy of visual prompting for adapting large-scale models in vision. As pixel space is inherently continuous, we follow

*Equal contribution.

* 平等贡献。

[2] For convenience, we unify the term and use ”visual prompt” to denote any pixel-space modification to the input image for model adaptation.

[2] 为了方便起见，我们统一术语，使用“视觉提示”来表示对输入图像进行的任何像素空间修改，以实现模型适应。

the recent approach that treats prompts as a continuous task-specific vector [12, 27, 14, 26]. We learn a single image perturbation (i.e., ”soft prompt”) via backpropagation while having the model parameters frozen. We map the model outputs to downstream labels by using a discrete text prompt for CLIP [37] and hard-coded mapping for vision models (Figure 1).

受到自然语言提示成功的启发，我们旨在研究视觉提示在适应大规模视觉模型中的有效性。由于像素空间本质上是连续的，我们遵循最近的方法，将提示视为一个连续的任务特定向量 [12, 27, 14, 26]。我们通过反向传播学习一个单一的图像扰动 (即“软提示”)，同时保持模型参数不变。我们通过使用 CLIP 的离散文本提示 [37] 和视觉模型的硬编码映射，将模型输出映射到下游标签 (图 1)。

How is visual prompting different from existing adaptation methods? Currently in vision, standard adaptation methods are fine-tuning and linear probe. Both approaches require some level of access to the model: entire parameters in the case of fine-tuning and model outputs (usually activations at the penultimate layer) in the case of linear probe. In contrast, visual prompting adapts the input to a model. After acquiring the visual prompt, it does not require model access at test time. This opens up unique applications [39]; input-space adaptation puts control in the hands of the end-user of the system. For instance, a pedestrian could wear a visual prompt that improves their visibility to cars, without having access to the car itself, nor its vision system.

视觉提示与现有适应方法有何不同？目前在视觉领域，标准的适应方法是微调和线性探测。这两种方法都需要对模型有一定程度的访问: 在微调的情况下需要访问整个参数，而在线性探测的情况下则需要访问模型输出 (通常是倒数第二层的激活)。相比之下，视觉提示则是对模型输入进行适应。在获取视觉提示后，测试时不需要访问模型。这为独特的应用打开了新的可能性 [39]；输入空间的适应将控制权交给了系统的最终用户。例如，行人可以佩戴一种视觉提示，以提高他们在汽车面前的可见性，而无需接触汽车本身或其视觉系统。

We conduct comprehensive experiments across four pre-trained models and 15 image classification datasets. We demonstrate that visual prompting is surprisingly effective for CLIP [37] and robust to distribution shift, achieving performance competitive with, and sometimes beyond, standard linear probes. We further analyze what properties of the downstream dataset, prompt design, and output transformation affect performance. Note that our goal is not to achieve the state-of-the-art performance on specific tasks, but instead to broadly explore a new paradigm for visual adaptation. The surprising effectiveness of visual prompting provides a new perspective on how to adapt and use pre-trained models in vision.

我们在四个预训练模型和 15 个图像分类数据集上进行了全面的实验。我们证明了视觉提示在 CLIP [37] 上出奇地有效，并且对分布变化具有鲁棒性，其性能与标准线性探测相当，有时甚至超越它。我们进一步分析了下游数据集的哪些属性、提示设计和输出转换会影响性能。请注意，我们的目标并不是在特定任务上实现最先进的性能，而是广泛探索视觉适应的新范式。视觉提示的惊人有效性为如何适应和使用预训练模型提供了新的视角。

2 Related Work

2 相关工作

2.1 Natural Language Prompting

2.1 自然语言提示

Our investigation is inspired by the recent success in natural language prompting. Prompting in NLP reformulates the downstream dataset into a (masked) language modeling problem, so that a frozen language model directly adapts to a new task without updating any parameters. A prompt consists of constructing a task-specific template (e.g., ”I felt so [MASK]”) and label words (e.g., ”happy/horrible”) to fill in the blank [13]. However, hand-crafting the right prompt requires domain expertise and a significant amount of effort.

我们的研究灵感来自于自然语言提示的近期成功。在自然语言处理中的提示将下游数据集重新表述为一个 (掩码) 语言建模问题，使得一个冻结的语言模型可以直接适应新任务，而无需更新任何参数。提示由构建特定任务的模板 (例如，“我感到如此 [MASK]”) 和填补空白的标签词 (例如，“快乐/可怕”) 组成 [13]。然而，手工制作合适的提示需要领域专业知识和大量的努力。

Prefix tuning [27] or prompt tuning [26] mitigates this problem by learning a ”soft prompt” via backpropagation, while having the model parameters fixed. Prefix tuning learns a task-specific continuous vector (i.e., prefix) that allows language models to adapt to various generation tasks. While prefix tuning prepends the prefix to each encoder layer, prompt tuning further simplifies by only prepending tunable tokens to the input. When applied to large models with billions of parameters, a properly optimized prompt achieves competitive performance to fine-tuning the entire model, while significantly reducing memory usage and per-task storage. As prompts in pixel space are inherently continuous, we follow this line of work and optimize the pixels directly.

前缀调优 [27] 或提示调优 [26] 通过反向传播学习“软提示”来缓解这个问题，同时保持模型参数不变。前缀调优学习一个特定于任务的连续向量 (即前缀)，使语言模型能够适应各种生成任务。前缀调优将前缀添加到每个编码器层，而提示调优则进一步简化，仅将可调节的标记添加到输入中。当应用于具有数十亿参数的大型模型时，经过适当优化的提示在性能上与微调整个模型相当，同时显著减少内存使用和每个任务的存储需求。由于像素空间中的提示本质上是连续的，我们遵循这一研究方向，直接优化像素。

2.2 Prompting with Images

2.2 使用图像进行提示

There have been initial approaches that attempt to prompt with images. Similar to prefix tuning, Frozen [44] creates a image-conditional prompt by training a vision encoder using gradients from a frozen language model. The images are represented as a continuous embedding from the vision encoder and used as a visual prefix to allow frozen language models to perform multi-modal tasks. CPT [46] converts visual grounding into a fill-in-the-blank problem by creating visual prompts with colored blocks and color-based textual prompts. However, both of these approaches focus on extending the capabilities of a language-based model. On the other hand, we focus on investigating the efficacy of prompting for visual representations and image classification datasets. In other words, we assume that the pre-trained model consists of a visual encoder and focus on reformulating image datasets. Visual prompt tuning [20] is concurrent work that proposes visual prompts specific to Vision Transformers [11]. It uses deep prompt tuning [27, 36, 29] by prepending a set of tunable parameters to each Transformer encoder layer.

已经有初步的方法尝试使用图像进行提示。与前缀调优类似，Frozen [44] 通过使用来自冻结语言模型的梯度训练视觉编码器来创建图像条件提示。图像被表示为来自视觉编码器的连续嵌入，并用作视觉前缀，以使冻结语言模型能够执行多模态任务。CPT [46] 通过创建带有彩色块和基于颜色的文本提示的视觉提示，将视觉定位转换为填空问题。然而，这两种方法都侧重于扩展基于语言的模型的能力。另一方面，我们专注于研究在视觉表示和图像分类数据集上提示的有效性。换句话说，我们假设预训练模型由视觉编码器组成，并专注于重新构建图像数据集。视觉提示调优 [20] 是一项同时进行的工作，提出了特定于视觉变换器 [11] 的视觉提示。它通过将一组可调参数添加到每个变换器编码器层，使用深度提示调优 [27, 36, 29]。

2.3 Adversarial Reprogramming and Unadversarial Examples

2.3 对抗性重编程与非对抗性示例

B Adversarial reprogramming [12] is a type of adversarial attack where a single, class-agnostic perturbation reprograms a model to perform a new task chosen by the attacker. Despite its adversarial goal, the framework essentially serves the same purpose as prompting: adapt a frozen model to new tasks by modifying the input and/or output of the downstream dataset. However, existing methods in vision [5, 22, 38] are designed to achieve an adversarial goal or demonstrate limited application to small-scale vision models and simple datasets. Similarly, unadversarial examples [39] aim to increase performance on the (pre-)trained task. It learns an image perturbation that improves performance on a specific class (i.e., class-conditional). In our work, we revisit adversarial reprogramming as a form of visual prompt and investigate its efficacy in adapting large-scale models in vision.

B 对抗性重编程 [12] 是一种对抗性攻击，其中单个与类别无关的扰动重新编程模型以执行攻击者选择的新任务。尽管其目标是对抗性的，但该框架本质上与提示的目的相同: 通过修改下游数据集的输入和/或输出，将一个冻结的模型适应于新任务。然而，现有的视觉方法 [5, 22, 38] 旨在实现对抗性目标，或仅在小规模视觉模型和简单数据集上展示有限的应用。同样，非对抗性示例 [39] 旨在提高在 (预) 训练任务上的性能。它学习一种图像扰动，以提高特定类别的性能 (即类别条件)。在我们的工作中，我们重新审视对抗性重编程作为一种视觉提示，并研究其在适应大规模视觉模型中的有效性。

2.4 Adapting Pre-trained Models in Vision

2.4 适应视觉中的预训练模型

Figure 2 provides a summary of different methods for adapting a pre-trained model. Fine-tuning and linear probing are highly flexible in their usage: they can be used to adapt the model to a new domain of inputs or to a new task with different output semantics. However, they also require some level of access to the model: parameters in the case of fine-tuning and model outputs (usually activations at the penultimate layer) in the case of linear probes. Domain adaptation is an interesting alternative to model adaptation in that it only modifies the inputs to the model using techniques such as image-to-image translation [50, 19]. Like domain adaptation, visual prompting also modifies the inputs to a model. Therefore, once the end user has found the visual prompt, it does not require having control over the model itself at test time. This opens up unique applications; for example, users can feed domain-adapted images to online APIs that can only be manipulated via their inputs. Domain adaptation focuses on adapting a source domain to look like a target domain, requiring both source and target datasets available at hand. On the other hand, we demonstfirate that visual prompting can steer model in more arbitrary ways; for example, a model that performs one classification task can be adapted to perform an entirely different classification task, with new output semantics, just by perturbing the input pixels. Also, whereas domain adaptation methods are typically input-conditional, the visual prompts we explore in this paper are fixed (i.e., input-agnostic) across an entire dataset, as in NLP where the same natural language prompt is added to all model queries.

图 2 提供了适应预训练模型的不同方法的总结。微调和线性探测在使用上非常灵活: 它们可以用于将模型适应于新的输入领域或具有不同输出语义的新任务。然而，它们也需要对模型有一定程度的访问: 微调的情况下需要访问参数，而线性探测则需要访问模型输出 (通常是倒数第二层的激活)。领域适应是模型适应的一个有趣替代方案，因为它仅通过图像到图像的转换等技术修改模型的输入 [50, 19]。与领域适应类似，视觉提示也修改模型的输入。因此，一旦最终用户找到了视觉提示，在测试时就不需要控制模型本身。这为独特的应用打开了可能性；例如，用户可以将领域适应的图像输入到只能通过其输入进行操作的在线 API 中。领域适应专注于将源领域调整为看起来像目标领域，要求手头有源数据集和目标数据集。另一方面，我们证明视觉提示可以以更任意的方式引导模型；例如，一个执行某一分类任务的模型可以通过扰动输入像素被调整为执行完全不同的分类任务，具有新的输出语义。此外，尽管领域适应方法通常是输入条件的，但我们在本文中探讨的视觉提示在整个数据集中是固定的 (即，与输入无关)，就像在自然语言处理 (NLP) 中，所有模型查询都添加相同的自然语言提示。

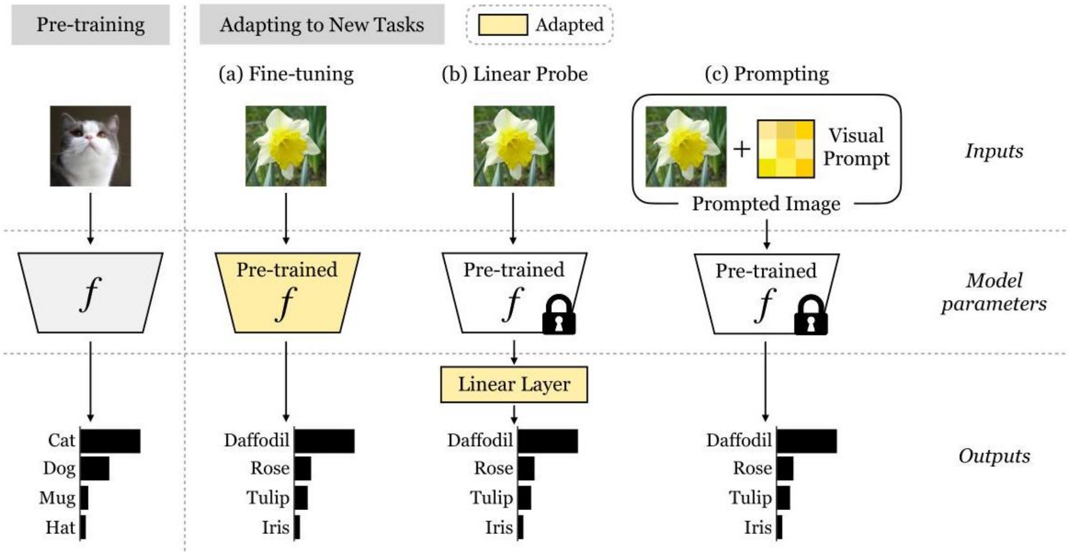

Figure 2: Methods for adapting pre-trained models to downstream tasks. (a) Fine-tuning adapts the entire model parameters. (b) Linear probes adapt the model outputs (usually activations at the penultimate layer) by learning a linear layer. (c) Prompting adapts the (downstream) dataset by reformulating the input and/or output.

图 2: 将预训练模型适应于下游任务的方法。(a) 微调适应整个模型参数。(b) 线性探针通过学习线性层来适应模型输出 (通常是倒数第二层的激活)。(c) 提示通过重新构造输入和/或输出来适应 (下游) 数据

集。

3 Methods

3 方法

Under different terms, prompting and adversarial reprogramming serve the same purpose: data-space adaptation. They generally consist of two stages [6, 28]: input transformation and output transformation. The goal of the input transformation (or prompt engineering) is to design a proper prompt that specifies the task which is applied to the input. The goal of the output transformation (or answer engineering) is to map the model's output/answer to the target label. We introduce different design choices for vision and vision-language models according to their pre-trained task.

在不同的术语下，提示和对抗性重编程服务于相同的目的: 数据空间适应。它们通常由两个阶段组成 [6, 28]: 输入转换和输出转换。输入转换 (或提示工程) 的目标是设计一个适当的提示，以指定应用于输入的任务。输出转换 (或答案工程) 的目标是将模型的输出/答案映射到目标标签。我们根据它们的预训练任务介绍不同的设计选择，以适应视觉和视觉-语言模型。

3.1 Pre-trained Models

3.1 预训练模型

Prompts in pixel form can essentially be applied to any visual representation. Therefore, we select three vision models and one vision-language model: Instagram-pretrained ResNeXt (Instagram) [32], Big Transfer (BiT-M) [24], ResNet trained on ImageNet-1k (RN50) [16, 10], and CLIP [37]. Vision models are trained to predict a fixed set of predetermined classes and typically require learning a separate layer to predict unseen classes. In contrast, CLIP is a vision-language model that is able to perform flexible zero-shot transfer to unseen classes using text prompts. We summarize the pre-trained model details in the Appendix. We select models across varying input modalities, pre-trained dataset size, and model architecture to evaluate the practical utility of visual prompts. For Instagram-pretrained ResNeXt, we use the model additionally fine-tuned on ImageNet-1k.

像素形式的提示本质上可以应用于任何视觉表示。因此，我们选择了三个视觉模型和一个视觉-语言模型:Instagram 预训练的 ResNeXt(Instagram) [32]，大迁移 (BiT-M) [24]，在 ImageNet-1k 上训练的 ResNet(RN50) [16, 10]，以及 CLIP [37]。视觉模型被训练以预测一组固定的预定类别，通常需要学习一个单独的层来预测未见类别。相比之下，CLIP 是一个视觉-语言模型，能够使用文本提示灵活地进行零样本迁移到未见类别。我们在附录中总结了预训练模型的详细信息。我们选择跨越不同输入模态、预训练数据集大小和模型架构的模型，以评估视觉提示的实际效用。对于 Instagram 预训练的 ResNeXt，我们使用了在 ImageNet-1k 上额外微调的模型。

3.2 Input Transformation

3.2 输入转换

There can be several ways of designing a visual prompt. As pixel space is less discrete compared to natural language, it is difficult to handcraft prompts as in NLP (e.g., "a photo of a [LABEL]" for image classification). In fact, it is unclear what type of visual context is useful for each downstream task (e.g., what visual information would be useful for specifying satellite image classification?). Intuitively, a visual prompt does not necessarily need to be interpretable to humans; it's a visual cue that aids the decision of a machine learning model. Thus, let the model optimize the visual context! We follow a simple gradient-based approach [12, 14, 27, 26] where we directly optimize the visual prompt via backpropagation.

设计视觉提示可以有多种方式。由于像素空间相比于自然语言不那么离散，因此很难像在自然语言处理 (NLP) 中那样手工制作提示 (例如，“一张 [LABEL] 的照片” 用于图像分类)。事实上，对于每个下游任务，什么类型的视觉上下文是有用的并不清楚 (例如，什么视觉信息对于指定卫星图像分类是有用的?)。直观上，视觉提示不一定需要对人类可解释；它是一个帮助机器学习模型决策的视觉线索。因此，让模型优化视觉上下文！我们遵循一种简单的基于梯度的方法 [12, 14, 27, 26]，通过反向传播直接优化视觉提示。

3.2.1 Prompt Tuning

3.2.1 提示调优

Given a frozen pre-trained model F and a downstream task dataset $\mathcal{D}=\left\{\left(x_1, y_1\right), \ldots,\left(x_m, y_m\right)\right\}$, our objective is to learn a single, task-specific visual prompt v_ϕ parameterized by ϕ . The prompt is added to the input image to form a prompted image $x+v_\phi$. During training, the model maximizes the likelihood of the correct label y ,

给定一个冻结的预训练模型 F 和一个下游任务数据集 $\mathcal{D}=\left\{\left(x_1, y_1\right), \ldots,\left(x_m, y_m\right)\right\}$ ，我们的目标是学习一个单一的、任务特定的视觉提示 v_ϕ ，其参数由 ϕ 指定。该提示被添加到输入图像中以形成一个提示图像 $x+v_\phi$ 。在训练过程中，模型最大化正确标签的似然性 y ，

$$\max_\phi P_{\theta ; \phi}\left(y \mid x+v_\phi\right) \tag{1}$$

while the gradient updates are applied only to the prompt parameters ϕ and the model parameters θ remain frozen. During evaluation, the optimized prompt is added to all test-time images,

同时梯度更新仅应用于提示参数 ϕ ，而模型参数 θ 保持冻结。在评估过程中，优化后的提示被添加到所有测试时图像中，

$$X_{\text {test }}=\left\{x_1+v_\phi, \ldots, x_n+v_\phi\right\} \tag{2}$$

which are then processed through the frozen model F .

然后通过冻结模型 F 进行处理。

Note that our goal is to explore visual prompts as a practical adaptation method. Therefore, we do not necessitate any adversarial constraint of making the perturbations imperceptible. Also, adversarial reprogramming assumes the downstream dataset to be lower-resolution than the pre-trained dataset, such that the input perturbation is padded around the downstream dataset. In real-world applications, the downstream dataset can have varying resolutions. Thus, we resize every dataset to the input size of the pre-trained model and add the prompt directly to the input region.

请注意，我们的目标是探索视觉提示作为一种实用的适应方法。因此，我们并不要求对扰动进行任何对抗性约束以使其不可察觉。此外，对抗性重编程假设下游数据集的分辨率低于预训练数据集，从而输入扰动在下游数据集周围进行填充。在实际应用中，下游数据集可能具有不同的分辨率。因此，我们将每个数据集调整为预训练模型的输入大小，并直接将提示添加到输入区域。

3.2.2 Prompt Design

3.2.2 提示设计

There are several ways to design a visual prompt in terms of template and size. We explore three visual templates: pixel patch at random location, pixel patch at fixed location, and padding. We explore various prompt sizes p , where the actual number of parameters is $C p^2$ for patches and $2 C p(H+W-2 p)$ for padding, where C, H, W are the image channels, height and width respectively. Section 6.2 shows that padding with $p=30$ achieves the best performance over other design choices. We use this as default for all our experiments.

在模板和大小方面，有几种方法可以设计视觉提示。我们探索了三种视觉模板: 随机位置的像素补丁、固定位置的像素补丁和填充。我们探索了各种提示大小 p ，其中补丁的实际参数数量为 $C p^2$ ，填充的参数数量为 $2 C p(H+W-2 p)$ ，其中 C, H, W 分别是图像通道、高度和宽度。第 6.2 节显示，使用 $p=30$ 进行填充在其他设计选择中表现最佳。我们将其作为所有实验的默认设置。

3.3 Output Transformation

3.3 输出转换

To map model outputs to the target label, we take a different approach for vision models and CLIP. Standard vision models treat image classes as a numeric id (e.g., "cat" is mapped to "index 1"). We use a hard-coded mapping [12] and arbitrarily map downstream class indices to pre-trained class indices, discarding unassigned indices for loss computation. For CLIP, a vision-language model, we utilize text prompts [37] as our output transformation function. Image classes are represented by text (e.g., ”cat”) which are then prompted (e.g., ”a photo of a [object]”) to specify context of the downstream task. Note that we use a single, fixed text prompt (see Appendix) and only optimize the visual prompt. We follow the protocol for CLIP zero-shot transfer and calculate cosine similarity of the embeddings for every class, which is normalized into a probability distribution via softmax. The class with the highest probability is chosen as the model output. The full overview for vision models and CLIP is illustrated in Figure 1.

为了将模型输出映射到目标标签，我们对视觉模型和 CLIP 采取了不同的方法。标准视觉模型将图像类别视为数字 ID(例如，“猫” 映射为 “索引 1”)。我们使用硬编码映射 [12]，并任意将下游类别索引映射到预训练类别索引，丢弃未分配的索引以进行损失计算。对于 CLIP，一个视觉语言模型，我们利用文本提示 [37] 作为我们的输出转换函数。图像类别通过文本表示 (例如，“猫”)，然后通过提示 (例如，“一张 [object] 的照片”) 来指定下游任务的上下文。请注意，我们使用一个固定的文本提示 (见附录)，并仅优化视觉提示。我们遵循 CLIP 零样本迁移的协议，并计算每个类别的嵌入的余弦相似度，该相似度通过 softmax 正规化为概率分布。具有最高概率的类别被选为模型输出。视觉模型和 CLIP 的完整概述如图 1 所示。

3.4 Implementation Details

3.4 实施细节

To learn the visual prompt, the objective function for CLIP is identical to its evaluation setting, i.e., we only compute cross entropy loss over images, where a set of prompted text strings is processed through the text encoder to produce weights of a linear classifier [37]. For vision models, we compute cross entropy loss over new class indices. For all experiments, we use the padding template with prompt size of 30 . All images are resized to 224 × 224 to match the input size of pre-trained models, and preprocessed identical to the evaluation setting of each model. We find that closely following the pre-trained model’s evaluation setting is important for learning a good prompt. All visual prompts are trained for 1,000 epochs. We use SGD with a learning rate of 40, which is decayed using cosine schedule [31]. We use a batch size of 256 for CLIP, 128 for BiT-M and RN50, and 32 for Instagram.

为了学习视觉提示，CLIP 的目标函数与其评估设置相同，即我们仅对图像计算交叉熵损失，其中一组提示文本字符串通过文本编码器处理以生成线性分类器的权重 [37]。对于视觉模型，我们对新的类别索引计算交叉熵损失。在所有实验中，我们使用提示大小为 30 的填充模板。所有图像都被调整为 224 × 224 以匹配预训练模型的输入大小，并且预处理与每个模型的评估设置相同。我们发现，紧密遵循预训练模型的评估设置对于学习良好的提示是重要的。所有视觉提示训练 1,000 个周期。我们使用学习率为 40 的 SGD，并使用余弦调度进行衰减 [31]。对于 CLIP，我们使用批量大小为 256，对于 BiT-M 和 RN50 为 128，对于 Instagram 为 32。

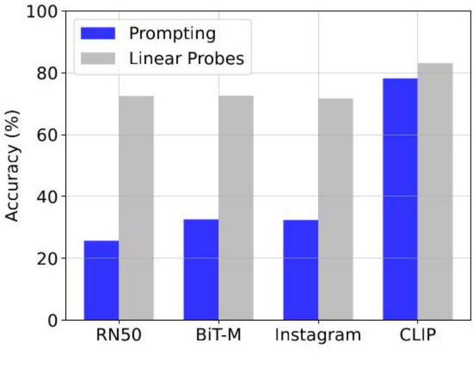

Figure 3: Prompting with vision models vs. CLIP. Prompting with vision models shows significant performance gap to linear probe. In contrast, prompting with CLIP achieves competitive performance.

图 3: 使用视觉模型与 CLIP 的提示。使用视觉模型的提示与线性探测器之间存在显著的性能差距。相比之下，使用 CLIP 的提示实现了具有竞争力的性能。

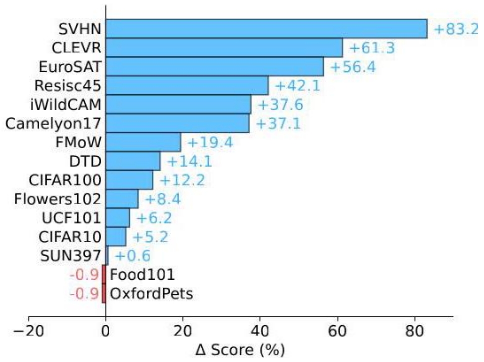

Figure 4: Accuracy gain from learning a visual prompt. The bars indicate the gain (or loss) in accuracy obtained by learning a single visual prompt compared to text-prompted (i.e., zero-shot) CLIP.

图 4: 通过学习视觉提示获得的准确性提升。条形图表示通过学习单个视觉提示与文本提示 (即零-shot)CLIP 相比所获得的准确性提升 (或损失)。

Table 1: Performance across 12 datasets using CLIP. TP, VP, LP, and FT refer to text prompt, visual prompt, linear probe, and fine-tuning respectively. The green shade indicates cases where visual prompting outperforms linear probe.

表 1: 使用 CLIP 在 12 个数据集上的性能。TP、VP、LP 和 FT 分别指文本提示、视觉提示、线性探测和微调。绿色阴影表示视觉提示优于线性探测的情况。

Model	Method	CIFAR100	CIFAR10	Flowers	Food	EuroSAT	SUN	UCF	SVHN	Pets	DTD	RESISC	CLEVR	Average
CLIP	TP	63.1	89.0	61.9	79.8	40.0	60.0	59.9	5.1	85.9	43.0	42.4	20.2	54.2
CLIP	VP + TP	75.3	94.2	70.3	78.9	96.4	60.6	66.1	88.4	85.0	57.1	84.5	81.4	78.2
CLIP	LP	80.0	95.0	96.9	84.6	95.3	75.0	83.3	65.4	89.2	74.6	92.3	66.0	83.1
CLIP	FT	82.1	95.8	97.4	80.5	97.9	64.0	80.9	95.7	88.5	72.3	93.3	94.4	86.9

模型	方法	CIFAR100	CIFAR10	花卉	食物	EuroSAT	SUN	UCF	SVHN	宠物	DTD	RESISC	CLEVR	平均
CLIP	TP	63.1	89.0	61.9	79.8	40.0	60.0	59.9	5.1	85.9	43.0	42.4	20.2	54.2
CLIP	VP + TP	75.3	94.2	70.3	78.9	96.4	60.6	66.1	88.4	85.0	57.1	84.5	81.4	78.2
CLIP	LP	80.0	95.0	96.9	84.6	95.3	75.0	83.3	65.4	89.2	74.6	92.3	66.0	83.1
CLIP	FT	82.1	95.8	97.4	80.5	97.9	64.0	80.9	95.7	88.5	72.3	93.3	94.4	86.9

4 Experimental Setup

4 实验设置

4.1 Datasets

4.1 数据集

To evaluate how well visual prompts adapt a model to new tasks, we measure performance across 12 datasets: CIFAR100, CIFAR10 [25], Flowers102 [34], Food101 [3], EuroSAT [17], SUN397 [45], DTD [9], UCF101 [43], SVHN [33], OxfordPets [35], Resisc45 [7], and CLEVR [21]. We also measure robustness to distribution shift, i.e., training distribution differs from the test distribution, by evaluating on three image classification datasets in WILDS [23]: Camelyon17 [1], FMoW [8], and iWildCAM [2]. For Camelyon17, training and test sets comprise tissue patches from different hospitals. For FMoW, training and test sets are from different regions and years. Finally, iWildCAM consists of photos from disjoint sets of camera traps. Note that we learn the visual prompt on the training set and evaluate its performance on the test set.

为了评估视觉提示如何使模型适应新任务，我们在 12 个数据集上测量性能:CIFAR100、CIFAR10 [25]、Flowers102 [34]、Food101 [3]、EuroSAT [17]、SUN397 [45]、DTD [9]、UCF101 [43]、SVHN [33]、OxfordPets [35]、Resisc45 [7] 和 CLEVR [21]。我们还通过在 WILDS [23] 中评估三个图像分类数据集来测量对分布变化的鲁棒性，即训练分布与测试分布不同:Camelyon17 [1]、FMoW [8] 和 iWildCAM [2]。对于 Camelyon17，训练集和测试集由来自不同医院的组织切片组成。对于 FMoW，训练集和测试集来自不同的地区和年份。最后，iWildCAM 由来自不重叠的相机陷阱集的照片组成。请注意，我们在训练集上学习视觉提示，并在测试集上评估其性能。

4.2 Baseline Methods

4.2 基线方法

To measure how visual prompting performs compared to existing adaptation methods (Figure 2), we compare fine-tuning, linear probes, and text prompting (i.e., zero-shot transfer). Fine-tuning and linear probe are standard adaptation methods in vision. Fine-tuning update the entire model parameters during adaptation. Linear probe is a lightweight alternative which adapts the model outputs (usually activations at the penultimate layer) by learning a linear layer, while having the model parameters frozen. For text prompting, we use ”This is a photo of a [LABEL]” as default. For CLEVR, we use ”This is a photo of [LABEL] objects”, with class label ”three” to ”ten”. For Camelyon17, we use ”a tissue region [LABEL] tumor”, with class label ”containing” and ”not containing”.

为了测量视觉提示与现有适应方法的性能比较 (图 2)，我们比较了微调、线性探针和文本提示 (即零样本迁移)。微调和线性探针是视觉领域的标准适应方法。微调在适应过程中更新整个模型参数。线性探针是一种轻量级替代方案，通过学习一个线性层来适应模型输出 (通常是倒数第二层的激活)，同时保持模型参数不变。对于文本提示，我们使用“这是一张 [LABEL] 的照片”作为默认提示。对于 CLEVR，我们使用“这是一张 [LABEL] 物体的照片”，类别标签为“三”到“十”。对于 Camelyon17，我们使用“一个组织区域 [LABEL] 肿瘤”，类别标签为“包含”和“不包含”。

5 Results

5 结果

5.1 Effectiveness of CLIP

5.1 CLIP 的有效性

We first compare prompting performance with linear probe, the current de facto approach to lightweight adaptation. Figure 3 shows average test accuracy across 12 datasets for each pre-trained model. Prompting with vision models, or adversarial reprogramming, shows significant performance gap (+40%) to standard linear probe. On the other hand, we find that prompting is surprisingly effective for CLIP, achieving competitive performance to linear probe. In particular, prompting outperforms linear probe on EuroSAT, SVHN, and CLEVR, by 1.1%, 23%, and 15.4% respectively (Table 1). On average, learning a visual prompt achieves 24% performance gain compared to using text prompt only (i.e., ”zero-shot transfer”). Interestingly, we find that the performance of visual prompts varies across datasets (Figure 4). Regarding this phenomenon, we further analyze what properties of the dataset affect performance in Section 6.1. We report full results across 12 datasets for vision models in the Appendix.

我们首先将提示性能与线性探测进行比较，线性探测是当前轻量级适应的事实标准方法。图 3 显示了每个预训练模型在 12 个数据集上的平均测试准确率。使用视觉模型的提示，或对抗性重编程，显示出与标准线性探测之间显著的性能差距 (+40%)。另一方面，我们发现提示对于 CLIP 出乎意料地有效，达到了与线性探测相竞争的性能。特别是，提示在 EuroSAT、SVHN 和 CLEVR 上分别超越了线性探测 1.1%、23% 和 15.4%(表 1)。平均而言，学习视觉提示相比仅使用文本提示 (即“零样本迁移”) 实现了 24% 的性能提升。有趣的是，我们发现视觉提示的性能在不同数据集之间有所不同 (图 4)。关于这一现象，我们在第 6.1 节进一步分析了数据集的哪些属性影响性能。我们在附录中报告了 12 个数据集上视觉模型的完整结果。

5.2 Robustness to Distribution Shift

5.2 对分布变化的鲁棒性

As model parameters remain frozen, prompting prevents modifying the general knowledge base of the pre-trained model. This reduces the possibility of overfitting to spurious correlations in the downstream dataset, thereby improving robustness to distribution shift. Using the WILDS benchmark [23], we learn visual prompts from training sets that contain images from a particular domain, and see how it transfers to test sets from different domains (e.g., images from different hospitals, regions, years, cameras). Table 2 show that average performance gap compared to linear probe and fine-tuning is further reduced to 4.5%

and 3.5% respectively. On Camelyon17, visual prompting outperforms both linear probe and fine-tuning by 4.9% and 6.5% respectively. This suggests the practical utility of prompting in real-world deployments, where diverse range of domain shifts naturally arise. We report robustness results for vision models in the Appendix.

当模型参数保持不变时，提示防止修改预训练模型的一般知识库。这减少了对下游数据集中虚假相关性的过拟合可能性，从而提高了对分布变化的鲁棒性。使用 WILDS 基准 [23]，我们从包含特定领域图像的训练集中学习视觉提示，并观察其如何转移到来自不同领域的测试集 (例如，来自不同医院、地区、年份、相机的图像)。表 2 显示与线性探测和微调相比，平均性能差距进一步减少到 4.5% 和 3.5% 。在 Camelyon17 上，视觉提示分别超越线性探测和微调 4.9% 和 6.5% 。这表明在实际部署中，提示具有实用价值，因为在这些情况下自然会出现多种领域变化。我们在附录中报告了视觉模型的鲁棒性结果。

Table 2: Out-of-distribution test accuracy using CLIP.

表 2: 使用 CLIP 的分布外测试准确性。

Model	Method	iWILDCAM	FMoW	Camelyon17	Average
CLIP	TP	14.1	13.5	52.7	26.8
CLIP	VP + TP	51.7	32.9	89.8	58.1
CLIP	LP	66.7	36.3	84.9	62.6
CLIP	FT	54.9	46.6	83.3	61.6

模型	方法	iWILDCAM	FMoW	Camelyon17	平均
CLIP	TP	14.1	13.5	52.7	26.8
CLIP	VP + TP	51.7	32.9	89.8	58.1
CLIP	LP	66.7	36.3	84.9	62.6
CLIP	FT	54.9	46.6	83.3	61.6

6 Understanding Visual Prompts

6 理解视觉提示

In this section, we investigate visual prompting performance in regard to properties of the downstream dataset, prompt design (i.e., input transformation), and output transformation.

在本节中，我们研究视觉提示性能与下游数据集的属性、提示设计 (即输入转换) 和输出转换之间的关系。

6.1 Downstream Dataset

6.1 下游数据集

We find that the performance of visual prompting varies across downstream datasets. As shown in Figure 4, the best-performing dataset achieves +83.2% accuracy gain, while the worst-performing dataset has -1% accuracy loss. To explain this phenomenon, we first hypothesize that visual prompts bridge the distribution gap by converting the unfamiliar downstream dataset to look more similar to the pre-trained dataset. Under this hypothesis, visual prompts would not help datasets already within the pre-trained distribution, yet could help datasets that are severely out-of-distribution. While CLIP's pre-trained dataset is not available to the public, it is excessively tuned to achieve state-of-the-art zero-shot performance on ImageNet. Thus, we use ImageNet as a proxy. We validate our hypothesis by measuring the distributional similarity between ImageNet and downstream datasets using the FID score [18]. We compare these scores to accuracy gain from visual prompts. Due to computation limitations, we randomly sample 100k images from the ImageNet-1k training set to compute the metrics. In Figure 5, we observe a general performance gain as the downstream dataset becomes more out-of-distribution to ImageNet (e.g., CLEVR, SVHN). Another hypothesis regards to learning a single prompt per dataset. While this may be sufficient for datasets with low perceptual diversity, a single visual prompt may fail to capture the full distribution as the diversity increases. We measure perceptual diversity using LPIPS [47]. For each dataset, we measure LPIPS between two randomly sampled image pairs and report the average score. Figure 5 shows that a learning single visual prompt achieves better performance gain for datasets with low perceptual diversity.

我们发现视觉提示的性能在下游数据集之间有所不同。如图 4 所示，表现最佳的数据集实现了 +83.2% 的准确率提升，而表现最差的数据集则有 −1% 的准确率损失。为了解释这一现象，我们首先假设视觉提示通过将不熟悉的下游数据集转换为看起来更类似于预训练数据集，从而弥合分布差距。在这一假设下，视觉提示对已经在预训练分布内的数据集没有帮助，但可以帮助那些严重超出分布的数据集。虽然 CLIP 的预训练数据集并未对公众开放，但它经过过度调优，以在 ImageNet 上实现最先进的零-shot 性能。因此，我们使用 ImageNet 作为代理。我们通过使用 FID 分数 [18] 测量 ImageNet 和下游数据集之间的分布相似性来验证我们的假设。我们将这些分数与视觉提示带来的准确率提升进行比较。由于计算限制，我们随机从 ImageNet-1k 训练集中抽取 100k 张图像来计算指标。在图 5 中，我们观察到随着下游数据集与 ImageNet 的分布差异增大 (例如 CLEVR、SVHN)，整体性能提升。另一个假设涉及每个数据集学习一个单一的提示。虽然这对于感知多样性低的数据集可能是足够的，但随着多样性的增加，单一的视觉提示可能无法捕捉到完整的分布。我们使用 LPIPS [47] 测量感知多样性。对于每个数据集，我们测量两个随机抽样的图像对之间的 LPIPS，并报告平均分数。图 5 显示，对于感知多样性低的数据集，学习单一视觉提示能够实现更好的性能提升。

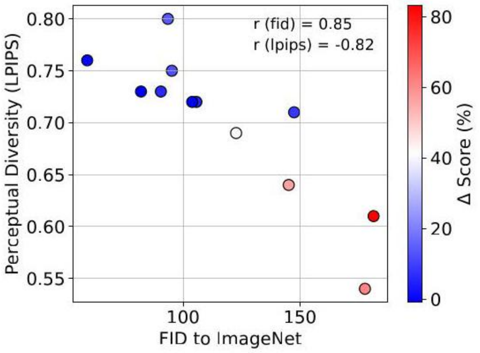

Figure 5: What properties of the downstream dataset affect performance? We see a general performance gain as (1) datasets become more out-of-distribution to ImageNet, and (2) perceptual diversity of a dataset decreases.

图 5: 下游数据集的哪些属性影响性能？我们观察到，随着 (1) 数据集与 ImageNet 的分布差异增大，以及 (2) 数据集的感知多样性降低，整体性能提升。

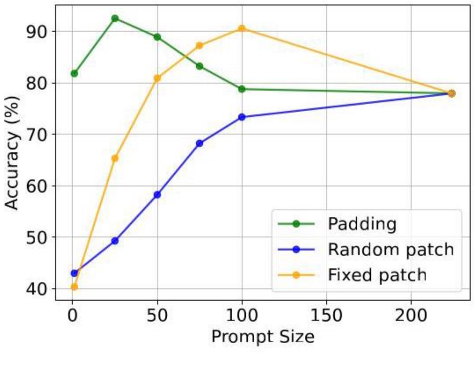

Figure 6: How does prompt design affect performance? Using a moderate-size, padding template achieves best results on image classification tasks.

图 6: 提示设计如何影响性能？使用中等大小的填充模板在图像分类任务中取得最佳结果。

6.2 Prompt Design

6.2 提示设计

Choosing the right prompt design (i.e., template and size) can highly affect performance. We perform an ablation study on three different templates: pixel patch at random location, pixel patch at fixed location, and padding, across prompt size $p = 1, \ldots, 224$. We measure accuracy on the EuroSAT dataset using a frozen CLIP. Figure 6.2 shows that using a fixed-location template (i.e., padding, fixed patch) yields better performance. For fixed-location templates, we find that performance improves as prompt size increases (i.e., more trainable parameters), then it starts to drop for +70k parameters. Surprisingly, we find that our simplest approach - adding a single-pixel prompt —can yield a 3% improvement over text-prompted CLIP (Figure 7). Overall, padding with $p = 30$ achieves the best performance in our experiments. We believe this is because our application scope is image classification, where the object of interest tends to be located in the center of the image. We believe other visual tasks may require significantly different design choices. Refer to Section 3.2.2 on how the actual number of parameters are calculated.

选择合适的提示设计 (即模板和大小) 可以极大地影响性能。我们对三种不同的模板进行了消融研究: 随机位置的像素补丁、固定位置的像素补丁和填充，跨提示大小 $p = 1, ..., 224$ 。我们使用冻结的 CLIP 在 EuroSAT 数据集上测量准确性。图 6.2 显示，使用固定位置模板 (即填充、固定补丁) 可以获得更好的性能。对于固定位置模板，我们发现随着提示大小的增加 (即更多可训练参数)，性能有所提升，然后在 +70k 参数时开始下降。令人惊讶的是，我们发现我们最简单的方法——添加一个单像素提示——可以比文本提示的 CLIP 提高 3% 的准确率 (图 7)。总体而言，使用 $p = 30$ 的填充在我们的实验中实现了最佳性能。我们认为这是因为我们的应用范围是图像分类，其中感兴趣的对象往往位于图像的中心。我们认为其他视觉任务可能需要显著不同的设计选择。有关实际参数数量计算的方法，请参见第 3.2.2 节。

Figure 7: Given a frozen CLIP, adding a single pixel to Eu-roSAT achieves +3% accuracy (zoom in to find the red pixel).

图 7: 给定一个冻结的 CLIP，向 EuroSAT 添加一个单像素可达到 +3% 的准确率 (放大以找到红色像素)。

6.3 Output Transformation

6.3 输出转换

We investigate prompting performance in regard to how we design the output transformation. For vision models, we follow [12] and use hard-coded mapping; downstream class indices are arbitrarily assigned to pre-trained class indices. We analyze how this mapping affects downstream performance. Using a subset of OxfordPets, we construct a simple toy dataset for classifying dogs and cats. Using ResNet trained on ImageNet-1k (RN50), we compare two cases: (1) downstream classes are assigned to pre-trained classes with similar semantics (unseen ”dog” assigned to pre-trained ”chihuahua” index),(2) we swap the indices (cat assigned to dog index, vice versa). (1) achieves 100% and (2) achieves 62.5% ; having similar semantics between class indices is critical for performance. This may explain the performance gap between vision models and CLIP. For CLIP, a vision-language model, we use text prompts for output transformation. As we learn visual prompts via backpropagation, the learning signal is dependent on the text prompt we use. It has been reported that CLIP's zero-shot accuracy can be significantly improved by using a better text prompt [37]. Therefore, we hypothesize that the quality of text prompt affects the performance of visual prompts. On EuroSAT, we measure text prompt quality by the zero-shot performance of CLIP. Figure 8 shows that the performance gain from visual prompts is higher for text prompts with low zero-shot performance. In other words, visual prompting can compensate for low-quality text prompts. As manually searching for the best text prompt is extremely laborsome, this result highlights the usefulness of visual prompts.

我们研究了输出转换设计对提示性能的影响。对于视觉模型，我们遵循 [12] 并使用硬编码映射；下游类别索引被任意分配给预训练类别索引。我们分析了这种映射如何影响下游性能。使用 OxfordPets 的一个子集，我们构建了一个简单的玩具数据集，用于分类狗和猫。使用在 ImageNet-1k (RN50) 上训练的 ResNet，我们比较了两种情况:(1) 下游类别被分配给具有相似语义的预训练类别 (未见的“狗”分配给预

训练的“吉娃娃”索引)，(2) 我们交换索引 (猫分配给狗索引，反之亦然)。(1) 达到 100% ，而 (2) 达到 62.5% ；类别索引之间具有相似语义对性能至关重要。这可能解释了视觉模型与 CLIP 之间的性能差距。对于 CLIP，一个视觉-语言模型，我们使用文本提示进行输出转换。当我们通过反向传播学习视觉提示时，学习信号依赖于我们使用的文本提示。有报道称，通过使用更好的文本提示，可以显著提高 CLIP 的零-shot 准确率 [37]。因此，我们假设文本提示的质量会影响视觉提示的性能。在 EuroSAT 上，我们通过 CLIP 的零-shot 性能来衡量文本提示的质量。图 8 显示，来自视觉提示的性能提升在低零-shot 性能的文本提示中更高。换句话说，视觉提示可以弥补低质量文本提示的不足。由于手动搜索最佳文本提示极为繁琐，这一结果突显了视觉提示的实用性。

7 Discussion

7 讨论

In this paper, we have investigated a method to perturb inputs to a pre-trained model in a manner which improves classification accuracy. A broader interpretation of visual prompting is to think of it as a way to steer a pre-trained model in any direction by modifying its input space. For instance, a visual prompt for an image-to-image model could be used to change the visual style of the input. Even though we have explored ”universal” visual prompts in this work (i.e., a single prompt that apply to all input images), prompts could also be made input-conditional and hence less universal but perhaps more accurate. The specific design choices including (a) input-specific or input-agnostic, (b) improving or decreasing accuracy, and (c) type of the pretrained-model, can be modified to create future interesting applications of prompting.

在本文中，我们研究了一种以改善分类准确性为目标的方式来扰动预训练模型的输入。对视觉提示的更广泛解释是将其视为通过修改输入空间来引导预训练模型朝任意方向发展的方式。例如，对于图像到图像模型的视觉提示可以用于改变输入的视觉风格。尽管我们在这项工作中探索了“通用”视觉提示 (即适用于所有输入图像的单一提示)，但提示也可以根据输入条件进行调整，因此可能不那么通用，但或许更为准确。具体的设计选择包括 (a) 输入特定或输入无关，(b) 提高或降低准确性，以及 (c) 预训练模型的类型，可以被修改以创造未来有趣的提示应用。

One natural question that arises following our exposition is in what situations would one prefer visual prompting over fine-tuning or a linear probe? Fine-tuning assumes that the model can be modified which may not always be the case (e.g., if the model is exposed by a API owned by a third-party). While prompting does under-perform linear probe in some cases, we would like to stress that the goal of this work is to show the existence of a prompting mechanism in ”pixel space”, which works across multiple datasets and pre-trained models, and reveals new avenues for how vision models can be effectively adapted. Our focus in this work is not to outperform state-of-the-art; we note that there are several approaches one could use to improve performance further including ensembling multiple prompts, using prompts in conjunction with linear probe or fine-tuning, or scaling the pre-trained model (e.g., ViT-L/14 of CLIP, which is unfortunately not available to the public). We leave these for future work.

在我们阐述之后，自然会产生一个问题: 在什么情况下人们会更倾向于使用视觉提示而不是微调或线性探测？微调假设模型可以被修改，但这并不总是成立 (例如，如果模型是由第三方拥有的 API 提供的)。虽然在某些情况下，提示的表现不如线性探测，但我们想强调的是，这项工作的目标是展示在“像素空间”中存在一种提示机制，该机制可以跨多个数据集和预训练模型工作，并揭示视觉模型如何有效适应的新途径。我们在这项工作中的重点并不是超越最先进的技术；我们注意到还有几种方法可以进一步提高性能，包括集成多个提示、将提示与线性探测或微调结合使用，或扩展预训练模型 (例如，CLIP 的 ViT-L/14，遗憾的是目前不对公众开放)。我们将这些留待未来的工作。

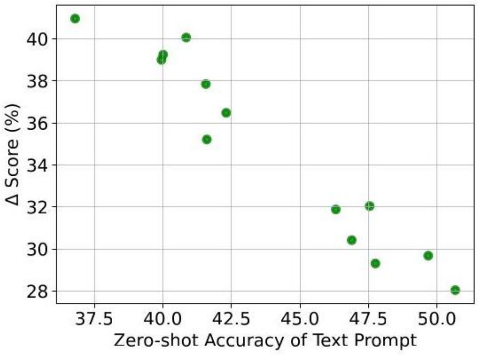

Figure 8: Visual prompts compensate for low-quality text prompts. We measure the quality of text prompts with the zero-shot accuracy of CLIP. As zero-shot accuracy decreases (i.e., lower quality text prompt), accuracy gain from visual prompting increases.

图 8: 视觉提示弥补低质量文本提示的不足。我们通过 CLIP 的零-shot 准确率来衡量文本提示的质量。随着零-shot 准确率的降低 (即，文本提示质量降低)，视觉提示带来的准确性提升增加。

8 Conclusion

8 结论

While standard adaptation methods in vision focus on introducing a separate task-specific head and adapt the model parameters or activations, we investigate visual prompting as practical adaptation method. We use a gradient-based scheme to learn a single, input-agnostic perturbation that repurposes a frozen model to perform a downstream task. Through various experiments across pre-trained models and datasets, we have demonstrated that CLIP is particularly suitable for visual prompting, achieving competitive results to linear probe. We hope that our unique findings will spur further research into: (1) better understanding pixel-space adaptation - when and why they are effective at steering deep networks, and (2) developing better visual prompts that further add to our repertoire of mechanisms for creating flexible and adaptable vision systems.

虽然视觉中的标准适应方法侧重于引入一个单独的任务特定头并调整模型参数或激活，但我们探讨了视觉提示作为一种实用的适应方法。我们使用基于梯度的方案来学习一个单一的、与输入无关的扰动，从而重新利用一个冻结的模型来执行下游任务。通过在预训练模型和数据集上的各种实验，我们证明了 CLIP 特别适合视觉提示，取得了与线性探测相竞争的结果。我们希望我们的独特发现能够激发进一步的研究: (1) 更好地理解像素空间适应——何时以及为什么它们在引导深度网络方面有效，以及 (2) 开发更好的视觉提示，进一步丰富我们创建灵活和可适应视觉系统的机制库。

Acknowledgements

致谢

We would like to thank Lucy Chai, Caroline Chan, Joanna Materzynska, Xavier Puig Fernandez, Minyoung Huh, Tongzhou Wang, and Yen-Chen Lin for proofreading the paper. We thank Judy Hoffman for helpful discussion and advice. This work was partially supported by funding from MIT STL and an MIT RSC award from the NEC fund.

我们要感谢 Lucy Chai、Caroline Chan、Joanna Materzynska、Xavier Puig Fernandez、Minyoung Huh、Tongzhou Wang 和 Yen-Chen Lin 对论文的校对。感谢 Judy Hoffman 的有益讨论和建议。本工作部分得到了 MIT STL 的资助和 NEC 基金的 MIT RSC 奖的支持。

References

参考文献

[1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Transactions on Medical Imaging, 2018.

[2] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. arXiv preprint arXiv:2004.10340, 2020.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101-mining discriminative components with random forests. In European conference on computer vision, pages 446-461. Springer, 2014.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877-1901, 2020.

[5] Lingwei Chen, Yujie Fan, and Yanfang Ye. Adversarial reprogramming of pretrained neural networks for fraud detection. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 2935-2939, 2021.

[6] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. arXiv preprint arXiv:2202.10629, 2022.

[7] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 105(10):1865-1883, 2017.

[8] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3606-3613, 2014.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248-255. Ieee, 2009.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[12] Gamaleldin F Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. International Conference on Learning Representations, 2019.

[13] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723, 2020.

[14] Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. Warp: Word-level adversarial reprogramming. arXiv preprint arXiv:2101.00121, 2021.

[15] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. arXiv preprint arXiv:2105.11259, 2021.

[16] 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.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, 和 Jian Sun. 深度残差学习用于图像识别. 见于 IEEE 计算机视觉与模式识别会议论文集, 第 770-778 页, 2016 年.

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(7):2217-2226, 2019.

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, 和 Damian Borth. Eurosat: 一种用于土地利用和土地覆盖分类的新数据集及深度学习基准. IEEE 应用地球观测与遥感选定主题期刊, 12(7):2217-2226, 2019.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, 和 Sepp Hochreiter. 通过两时间尺度更新规则训练的 GAN 收敛到局部纳什均衡. 神经信息处理系统进展, 30, 2017.

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In International conference on machine learning, pages 1989-1998. PMLR, 2018.

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, 和 Trevor Darrell. Cycada: 循环一致的对抗性领域适应. 在国际机器学习会议上, 页码 1989-1998. PMLR, 2018.

[20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. arXiv preprint arXiv:2203.12119, 2022.

[20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, 和 Ser-Nam Lim. 视觉提示调优. arXiv 预印本 arXiv:2203.12119, 2022.

[21] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2901-2910, 2017.

[21] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, 和 Ross Girshick. Clevr: 一种用于组合语言和基础视觉推理的诊断数据集. 在 IEEE 计算机视觉与模式

识别会议论文集中, 页码 2901-2910, 2017.

[22] Eliska Kloberdanz, Jin Tian, and Wei Le. An improved (adversarial) reprogramming technique for neural networks. In International Conference on Artificial Neural Networks, pages 3-15. Springer, 2021.

[22] Eliska Kloberdanz, Jin Tian, 和 Wei Le. 一种改进的 (对抗性) 神经网络重编程技术. 在国际人工神经网络会议上, 页码 3-15. Springer, 2021.

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In International Conference on Machine Learning (ICML), 2021.

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, 和 Percy Liang. WILDS: 一种野外分布转变的基准测试. 发表在国际机器学习会议 (ICML), 2021.

[24] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V 16, pages 491-507. Springer, 2020.

[24] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, 和 Neil Houlsby. 大迁移 (bit): 通用视觉表示学习. 发表在计算机视觉-ECCV 2020: 第 16 届欧洲会议, 英国格拉斯哥, 2020 年 8 月 23 日至 28 日, 会议论文集, 第五部分 16, 第 491-507 页. Springer, 2020.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.

[25] Alex Krizhevsky, Geoffrey Hinton 等. 从微小图像中学习多层特征.2009.

[26] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.

[26] Brian Lester, Rami Al-Rfou, 和 Noah Constant. 参数高效提示调优的规模效应. 发表在 2021 年自然语言处理实证方法会议 (EMNLP) 的会议论文集, 2021.

[27] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.

[27] Xiang Lisa Li 和 Percy Liang. 前缀调优: 优化生成的连续提示. arXiv 预印本 arXiv:2101.00190, 2021.

[28] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586, 2021.

[28] 彭飞刘、元伟哲、傅金兰、姜正宝、林宏明和格雷厄姆 · 纽比格。预训练、提示和预测: 自然语言处理中的提示方法系统性调查。arXiv 预印本 arXiv:2107.13586，2021 年。

[29] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602, 2021.

[29] 刘晓、季凯轩、傅怡成、杜正霄、杨志林和唐杰。P-tuning v2: 提示调优在各个规模和任务中可以与微调相媲美。arXiv 预印本 arXiv:2110.07602，2021 年。

[30] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. arXiv preprint arXiv:2103.10385, 2021.

[30] 刘晓、郑雅楠、杜正霄、丁铭、钱宇杰、杨志林和唐杰。GPT 也能理解。arXiv 预印本 arXiv:2103.10385，2021 年。

[31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

[31] 伊利亚 · 洛希奇洛夫和弗兰克 · 胡特。SGDR: 带有热重启的随机梯度下降。arXiv 预印本 arXiv:1608.03983，2016 年。

[32] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In Proceedings of the European conference on computer vision (ECCV), pages 181-196, 2018.

[33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pages 722-729. IEEE, 2008.

[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In 2012 IEEE conference on computer vision and pattern recognition, pages 3498-3505. IEEE, 2012.

[36] Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. arXiv preprint arXiv:2104.06599, 2021.

[37] 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 International Conference on Machine Learning, pages 8748-8763. PMLR, 2021.

[38] Ettore Randazzo, Alexander Mordvintsev, Eyvind Niklasson, and Michael Levin. Adversarial reprogramming of neural cellular automata. Distill, 6(5):e00027-004, 2021.

[39] Hadi Salman, Andrew Ilyas, Logan Engstrom, Sai Vemprala, Aleksander Madry, and Ashish Kapoor. Unadversarial examples: Designing objects for robust vision. Advances in Neural Information Processing Systems, 34, 2021.

[40] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676, 2020.

[41] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. arXiv preprint arXiv:2009.07118, 2020.

[42] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980, 2020.

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[44] Maria Tsimpoukelli, Jacob 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, 2021.

[45] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3485-3492, June 2010.

[46] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797, 2021.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586-595, 2018.

[48] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. arXiv preprint arXiv:2104.05240, 2021.

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134, 2021.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223-2232, 2017.

A Appendix

A 附录

Table 3: Overview of pre-trained models.

表 3: 预训练模型概述。

Model	Architecture	Modality	Pre-trained Dataset	Objective
CLIP [37]	ViT-B/32	Vision-language	400M image-text pairs	Contrastive
Instagram [32]	ResNext101-32x8d	Vision	3.5B Instagram photos	Cross Entropy
BiT-M 24	ResNet-50	Vision	14M ImageNet-21k	Cross Entropy
RN50 [16]	ResNet-50	Vision	1.2M ImageNet-1k	Cross Entropy

模型	架构	模态	预训练数据集	目标
CLIP [37]	ViT-B/32	视觉-语言	4 亿对图像-文本	对比
Instagram [32]	ResNext101-32x8d	视觉	35 亿 Instagram 照片	交叉熵
BiT-M 24	ResNet-50	视觉	1400 万 ImageNet-21k	交叉熵
RN50 [16]	ResNet-50	视觉	120 万 ImageNet-1k	交叉熵

Table 4: Performance across 12 datasets using vision pre-trained models. TP, VP, LP, and FT refer to text prompt, visual prompt, linear probe, and fine-tuning respectively. The green shade indicates cases where visual prompting outperforms linear probe.

表 4: 使用视觉预训练模型在 12 个数据集上的性能。TP、VP、LP 和 FT 分别指文本提示、视觉提示、线性探测和微调。绿色阴影表示视觉提示优于线性探测的情况。

Model	Method	CIFAR100	CIFAR10	Flowers	Food	EuroSAT	SUN	UCF	SVHN	Pets	DTD	RESISC	CLEVR	Average
Instagram	VP	16.7	62.1	22.9	9.9	85.4	2.2	15.4	53.8	18.6	29.1	41.4	30.9	32.4
Instagram	LP	64.0	90.1	92.7	65.8	90.6	58.1	76.6	48.0	94.5	70.9	79.2	30.2	71.7
Instagram	FT	77.8	77.8	94.5	75.6	97.4	56.7	72.9	96.8	93.9	73.5	93.4	87.9	83.2
BiT-M	VP	16.2	53.6	29.2	11.6	72.5	2.6	16.6	56.3	24.7	30.0	43.0	34.8	32.6
BiT-M	LP	73.0	90.8	99.1	72.6	94.4	49.5	72.9	49.8	85.2	68.4	87.7	28.4	72.6
BiT-M	FT	76.2	94.1	99.4	75.6	98.3	52.7	81.3	97.2	89.0	69.3	93.6	88.2	84.6
RN50	VP	10.1	54.5	14.0	5.1	78.7	1.1	9.5	57.1	10.8	8.2	29.9	29.5	25.7
RN50	LP	67.7	87.7	92.7	62.5	94.5	57.5	69.4	60.3	91.1	66.7	87.1	32.6	72.5
RN50	FT	79.9	94.1	96.9	73.2	96.5	55.9	76.7	96.9	92.3	66.7	93.4	89.3	84.3

模型	方法	CIFAR100	CIFAR10	花卉	食物	EuroSAT	SUN	UCF	SVHN	宠物	DTD	RESISC	CLEVR	平均
Instagram	副总裁	16.7	62.1	22.9	9.9	85.4	2.2	15.4	53.8	18.6	29.1	41.4	30.9	32.4
Instagram	LP	64.0	90.1	92.7	65.8	90.6	58.1	76.6	48.0	94.5	70.9	79.2	30.2	71.7
Instagram	FT	77.8	77.8	94.5	75.6	97.4	56.7	72.9	96.8	93.9	73.5	93.4	87.9	83.2
BiT-M	副总裁	16.2	53.6	29.2	11.6	72.5	2.6	16.6	56.3	24.7	30.0	43.0	34.8	32.6
BiT-M	LP	73.0	90.8	99.1	72.6	94.4	49.5	72.9	49.8	85.2	68.4	87.7	28.4	72.6
BiT-M	FT	76.2	94.1	99.4	75.6	98.3	52.7	81.3	97.2	89.0	69.3	93.6	88.2	84.6
RN50	副总裁	10.1	54.5	14.0	5.1	78.7	1.1	9.5	57.1	10.8	8.2	29.9	29.5	25.7
RN50	LP	67.7	87.7	92.7	62.5	94.5	57.5	69.4	60.3	91.1	66.7	87.1	32.6	72.5
RN50	FT	79.9	94.1	96.9	73.2	96.5	55.9	76.7	96.9	92.3	66.7	93.4	89.3	84.3

Table 5: Out-of-distribution test accuracy on WILDS using vision pre-trained models. The green shade indicates a case where visual prompting outperforms linear probe and fine-tuning.

表 5: 使用视觉预训练模型在 WILDS 上的分布外测试准确率。绿色阴影表示视觉提示优于线性探测和微调的情况。

Model	Method	iWILDCAM	FMoW	Camelyon17	Average
Instagram	VP	52.2	14.0	87.1	51.1
Instagram	LP	64.1	22.7	77.4	54.7
Instagram	FT	62.6	46.8	86.5	65.3
BiT-M	VP	48.7	15.7	83.3	49.2
BiT-M	LP	55.9	22.3	89.0	55.7
BiT-M	FT	59.0	49.8	81.8	63.5
RN50	VP	51.9	12.6	84.5	49.7
RN50	LP	62.7	28.7	90.2	60.5
RN50	FT	62.2	46.7	88.0	65.6

模型	方法	iWILDCAM	FMoW	Camelyon17	平均
Instagram	VP	52.2	14.0	87.1	51.1
Instagram	LP	64.1	22.7	77.4	54.7
Instagram	FT	62.6	46.8	86.5	65.3
BiT-M	VP	48.7	15.7	83.3	49.2
BiT-M	LP	55.9	22.3	89.0	55.7
BiT-M	FT	59.0	49.8	81.8	63.5
RN50	VP	51.9	12.6	84.5	49.7
RN50	LP	62.7	28.7	90.2	60.5
RN50	FT	62.2	46.7	88.0	65.6

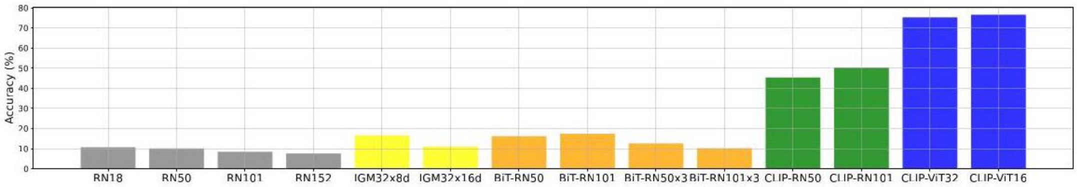

Figure 9: Model architecture ablation on CIFAR100.
图 9:CIFAR100 上的模型架构消融。

A.1 Ablation on Model Architecture

A.1 模型架构消融

In Figure 9, we compare performance using different model architectures on CIFAR100. We compare the original ImageNet-pretrained ResNets released by [16], namely ResNet-18, ResNet-50, ResNet- 101, ResNet-152. For Instagram-pre-trained ResNeXt [32], we compare two models (32x8d, 32x16d). For Big Transfer [24], we use four BiT-M models (ResNet-50, ResNet-101, ResNet-50x3, ResNet- 101x3). For ResNet-based CLIP models, we compare two models trained on 224 × 224 images (ResNet-50, ResNet-101). For CLIP models that use the Vision Transformer [11], we compare the two released models (ViT-B/32, ViT-B/16). For vision models, performance does not necessarily increase for larger models. For CLIP, we observe superiority of ViT-based models over ResNet-based models.

在图 9 中，我们比较了在 CIFAR100 上使用不同模型架构的性能。我们比较了 [16] 发布的原始 ImageNet 预训练 ResNets，即 ResNet-18、ResNet-50、ResNet-101 和 ResNet-152。对于 Instagram 预训练的 ResNeXt [32]，我们比较了两个模型 (32x8d，32x16d)。对于 Big Transfer [24]，我们使用了四个 BiT-M 模型 (ResNet-50、ResNet-101、ResNet-50x3、ResNet- 101x3)。对于基于 ResNet 的 CLIP 模型，我们比较了两个在 224 × 224 图像上训练的模型 (ResNet-50、ResNet-101)。对于使用视觉变换器的 CLIP 模型 [11]，我们比较了两个发布的模型 (ViT-B/32，ViT-B/16)。对于视觉模型，性能并不一定随着模型的增大而提高。对于 CLIP，我们观察到基于 ViT 的模型优于基于 ResNet 的模型。

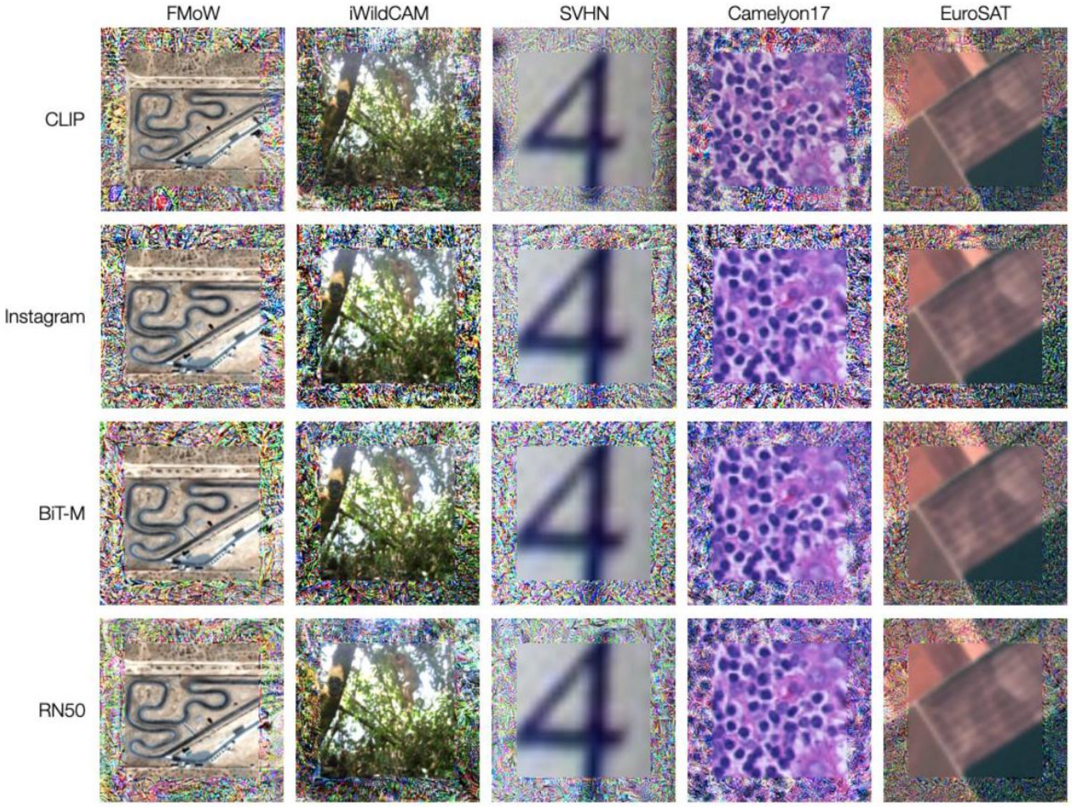

Figure 10: Visualizing task-specific visual prompts for each pre-trained model.
图 10: 为每个预训练模型可视化任务特定的视觉提示。
Table 6: Description of the datasets and the corresponding text prompt used for CLIP.
表 6: 数据集的描述及用于 CLIP 的相应文本提示。

Dataset	Train Size	Validation Size	Test Size	Classes	Text Prompt
CIFAR100	50,000	-	10,000	100	”This is a photo of a { }”
CIFAR10	50,000	-	10,000	10	”This is a photo of a { }”
Flowers102	4,093	1,633	2,463	102	”This is a photo of a { }”
Food101	50,500	20,200	30,300	101	”This is a photo of a { }”
EuroSAT	13,500	5,400	8,100	10	”This is a photo of a { }”
SUN397	15,888	3,970	19,850	397	”This is a photo of a { }”
UCF101	7,639	1,898	3,783	101	”This is a photo of a { }”
SVHN	73,257	-	26,032	10	”This is a photo of a { }”
OxfordPets	2,944	736	3,669	37	”This is a photo of a { }”
DTD	2,820	1,128	1,692	47	”This is a photo of a { }”
Resisc45	18,900	6,300	6,300	45	”This is a photo of a { }”
CLEVR/count	70,000	-	15,000	8	”This is a photo of { } objects”
iWildCAM	129,809	14,961	42,791	182	”This is a photo of a { }”
FMoW	76,863	19,915	22,108	62	”This is a photo of a { }”
Camelyon17	302,436	34,904	85,054	2	”a tissue region { } tumor”

数据集	训练规模	验证规模	测试大小	类别	文本提示
CIFAR100	50,000	-	10,000	100	”这是一张 { } 的照片”
CIFAR10	50,000	-	10,000	10	”这是一张 { } 的照片”
Flowers102	4,093	1,633	2,463	102	”这是一张 { } 的照片”
Food101	50,500	20,200	30,300	101	”这是一张 { } 的照片”
EuroSAT	13,500	5,400	8,100	10	”这是一张 { } 的照片”
SUN397	15,888	3,970	19,850	397	”这是一张 { } 的照片”
UCF101	7,639	1,898	3,783	101	”这是一张 { } 的照片”
SVHN	73,257	-	26,032	10	”这是一张 { } 的照片”
OxfordPets	2,944	736	3,669	37	”这是一张 { } 的照片”
DTD	2,820	1,128	1,692	47	”这是一张 { } 的照片”
Resisc45	18,900	6,300	6,300	45	”这是一张 { } 的照片”
CLEVR/count	70,000	-	15,000	8	”这是一个 { } 物体的照片”
iWildCAM	129,809	14,961	42,791	182	”这是一张 { } 的照片”
FMoW	76,863	19,915	22,108	62	”这是一张 { } 的照片”
Camelyon17	302,436	34,904	85,054	2	”一个组织区域 { } 肿瘤”

A.2 Dataset Statistics

A.2 数据集统计

Table 6 illustrates description of the datasets and the corresponding text prompt used for adapting CLIP. For OxfordPets, Flowers102, Food101, SUN397, DTD, EuroSAT, and UCF101, we used the data splits provided by [49]. For other datasets, we used the officially provided data splits.

表 6 展示了数据集的描述及用于适应 CLIP 的相应文本提示。对于 OxfordPets、Flowers102、Food101、SUN397、DTD、EuroSAT 和 UCF101，我们使用了 [49] 提供的数据划分。对于其他数据集，我们使用了官方提供的数据划分。

A.3 Change Log

A.3 更新日志

ArXiv v2 In ArXiv v1, we overlooked the adversarial reprogramming literature. In fact, visual prompting for vision models is essentially the same as adversarial reprogramming! In the current version, we have clarified this and removed claims of methodological novelty. We have reframed the paper as an exploration of the viability of visual prompts as a practical adaptation method for modern large-scale models. We thank Seong Joon Oh and users on twitter for pointing out the connection to adversarial reprogramming.

ArXiv v2 在 ArXiv v1 中，我们忽视了对抗性重编程文献。实际上，视觉模型的视觉提示本质上与对抗性重编程是相同的！在当前版本中，我们对此进行了澄清，并删除了方法新颖性的声明。我们将论文重新框定为对视觉提示作为现代大规模模型的实用适应方法的可行性探索。我们感谢 Seong Joon Oh 和 Twitter 上的用户指出了与对抗性重编程的联系。