Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners

可微分提示使预训练的语言模型更好

# Abstract

大规模预训练语言模型通过展示作为少样本学习者的卓越能力，对自然语言处理做出了重大贡献。然而，它们的有效性主要取决于模型参数的缩放和快速设计，阻碍了它们在大多数实际应用中的实现。本研究提出了一种名为 DifferentiAble pRompT (DART) 的新型可插拔、可扩展和高效方法，该方法可以将小型语言模型转换为更好的少样本学习器，而无需任何提示工程。这种方法背后的主要原理涉及将潜在的自然语言处理任务重新构建为预训练语言模型的任务，并通过反向传播对提示模板和目标标签进行差异化优化。此外，所提出的方法可以是：（i）插入任何预先训练的语言模型； (ii) 扩展到广泛的分类任务。对标准 NLP 任务的综合评估表明，所提出的方法实现了更好的少拍性能。

# 1 Introduction

预训练-微调范式已成为自然语言处理 (NLP) 的事实上的标准，并在多个基准测试中取得了优异的成绩 [10, 27, 22, 13, 3]。这些先驱者的成功似乎表明，大规模预训练模型始终是提升机器智能的灵丹妙药。然而，监督微调在实践中仍然容易出现标记数据，并且由于领域、语言和任务的变化而面临不可忽视的挑战。这些缺点导致了一项重要技术的研究，即小样本学习，它可以通过仅访问少量标记示例来显着提高机器智能和实际自适应应用的学习能力。

[6] 引入的 GPT-3 模型展示了令人印象深刻的小样本学习能力。给定自然语言提示和 16 个标记样本作为上下文输入中的演示，GPT-3 实现了 80% 的 SOTA 结果。然而，GPT-3 是一个具有 175B 参数的完全密集的变压器模型，这使得在大多数实际应用中部署具有挑战性。

最近，出现了一种新兴的微调方法，可以为较小的语言模型 (LM) 配备少量功能：通过完成完形填空任务 [39, 38, 14, 26] 将预训练的 LM 直接用作预测器，它将下游任务视为（屏蔽）语言建模问题。这些提示可用于微调，为分类器提供额外的任务信息，尤其是在低数据情况下。值得注意的是，Scao 等人。 [37] 观察到提示通常可以在多个分类任务中平均补偿数百个数据点。然而，确定适当的提示需要领域专业知识，而手工制作高性能提示通常需要不切实际的大验证集 [32]。最近的研究 [28, 51] 报告说，手动提示格式可能是次优的，这将导致准确性从随机猜测性能变化到接近最新技术水平。因此，以前的方法试图自动搜索离散的提示标记。然而，对于广泛的分类任务来说，获得优化的提示模板和目标标签标记并非易事。例如，特定的分类任务，例如带有alternate\_name 和country\_of \_birth 标签的关系提取，无法在词汇表中指定单个标签标记。

在本文中，我们提出了一种新颖的可区分提示 (DART) 微调方法，该方法与模型无关、参数高效且无需即时工程。如图 1 所示，关键思想是利用语言模型中的一些参数（未使用的标记）作为模板和标签标记，并使用反向传播在连续空间中优化它们。随后，我们引入了可微分提示学习以获得优化的提示模板和标签。由于有限样本的微调会受到不稳定性的影响 [12, 50]，我们提出了一种两阶段优化算法，首先学习模板和标签，然后是整体参数。我们进一步引入了一个辅助流畅性约束对象来确保提示嵌入之间的关联。

我们对 15 个 NLP 数据集进行了大量实验。在所有任务中只有少数训练样本，我们的方法 (DART) 可以获得更好的性能。值得注意的是，在具有复杂标签语义的关系提取数据集上，在 K = 8（完全监督设置为 1.55%）的设置中，与传统微调相比，绝对性能提升高达 23.28%。我们的方法可以应用于现实世界的分类任务，而无需收集和注释大量数据的高成本。本研究的主要贡献如下：

• 我们为小样本学习提出了一个新的简单框架，它是可插拔的、可扩展的和高效的，无需及时工程。据我们所知，在连续空间中优化标签标记也是语言模型提示中尚未探索的新研究分支。

• 对 15 个 NLP 任务的系统评估表明，这种简单而有效的方法有助于改进所有这些任务。值得注意的是，给定每类只有 8 个标记样本，我们提出的方法可以实现 SOTA 结果（完整数据集）的 90% 性能。

# 2 Related Work

语言模型提示。语言模型提示随着 GPT-3 [6] 的引入而出现，它展示了出色的少拍性能 [25]。然而，GPT-3 并不是为微调而设计的；它主要依赖于手工提示（上下文学习 [24, 51, 11, 29]）。因此，最近在该领域进行的研究 [34, 15, 7] 一直专注于自动搜索提示。希克等人。 [39, 38] 提出了 PET，它将 NLP 任务重新制定为完形填空式问题并执行基于梯度的微调。谭等人。 [42] 在微调期间使用更密集的监督对象改进 PET。申等人。 [40] 建议使用 AUTOPROMPT 为基于梯度引导搜索的各种任务创建提示。韩等人。 [17] 提出了一种称为 PTR 的方法，它利用逻辑规则为多类文本分类构造带有子提示的提示。王等人。 [44] 将潜在的 NLP 任务重新表述为一个蕴涵任务，然后用少量样本对模型进行微调。胡等人。 [19] 提出了一种通过校准将外部知识图整合到语言表达器中的方法。此外，高等人。 [14] 提出了 LM-BFF——更好的语言模型的少量微调，它利用 T5 [35] 在词汇表中生成模板和搜索标签标记。然而，使用生成模型和带有验证的标签搜索是计算密集型的。此外，由于神经网络的连续性，对离散空间的快速搜索是次优的。

为了克服这些限制，Liu 等人。 [26] 提出了 P-tuning，它采用了由 LSTM 学习的可训练的连续提示嵌入。钟等人。 [52] 提出了一种称为 OPTIPROMPT 的有效连续方法来优化事实探查的提示。李等人。 [26] 提出前缀调整，它保持语言模型参数冻结，但为自然语言生成任务优化了一个小的连续任务特定向量。莱斯特等人。 [21] 提出了一种学习“软提示”的机制，以调节冻结语言模型以执行下游任务。然而，这些方法仍然需要优化外部参数（例如 P-tuning 中的 LSTM）并且容易出现复杂的标签空间。

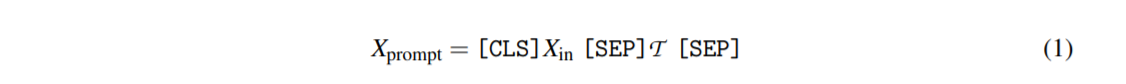
相反，本研究旨在开发一种基于不需要即时工程（包括模板和标签）和外部参数优化的预训练语言模型的新型小样本学习框架。此外，所提出的方法仅利用模型的非侵入性修改，可以插入任何预先训练的语言模型并扩展到广泛的分类任务。

少镜头学习。小样本学习可以通过仅访问少量标记示例 [49] 来显着提高机器智能和实际自适应应用的学习能力。所提出的方法对应于其他小样本 NLP 方法，包括：（1）元学习 [48, 4, 2, 9, 8, 47]，其中优化了辅助任务的数量。 (2) 中级训练 [33, 46]，通过对数据丰富的监督任务的进一步训练来补充预训练的 LM。 (3) 半监督学习 [30, 45]，利用未标记的样本。所提出的方法侧重于更现实的小样本设置（每个类的标记实例数量可以是任何变量）。

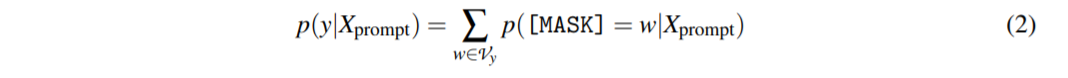
# 3 Background

## 3.1 Language Model Prompting

令 Xin = {x1, x2,..., xL} 是一个句子，其中 xi 是输入句子中的第 i 个标记，L 是标记的数量。 具体来说，Xin 被转换为一个固定的标记序列 X~in，然后映射到一个隐藏向量序列 {hk ∈ R d}。 给定输入序列 X~in = [CLS]Xin[SEP]，传统的微调方法利用 [CLS] 嵌入（例如，MLP 层）上的通用头层来预测输出类。 对于基于提示的方法，特定任务的模式字符串（模板 T ）旨在诱导模型生成对应于给定类（标签标记 M (Y)）的文本输出——我们将这两件事一起称为 一个提示。 具体来说，包含一个 [MASK] 令牌的 Xprompt 直接将 MLM 输入分配给以下任务：



当提示输入MLM时，模型可以得到候选类的概率分布p([MASK]|(Xprompt)，y∈Y为：



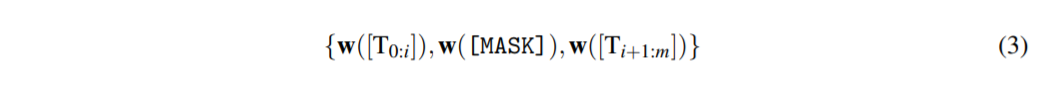
其中 w 表示第 y 类的第 w 个标签标记。 为了进一步理解语言模型提示的机制，我们从理论上分析了潜在的直觉。

# 4 Our Approach

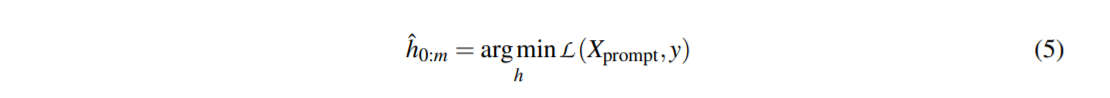
## 4.1 Motivation

从之前的经验发现 [14, 37] 可以看出，对于少样本学习者的预训练语言模型的改进，需要最佳提示。 由于具有离散标记的模板可能是次优的并且不足以表示特定的类 3，因此本研究提出了可区分的 pRompT，简称 DART，它可以减少即时工程的要求，以提高所提出的方法在各种情况下的适用性。 域。

## 4.2 Differentiable Template Optimization

由于语言标记是离散变量，因此通过标记搜索找到最佳提示并非易事，并且很容易陷入局部最小值。 为了克服这些限制，我们利用伪标记来构建模板，然后通过反向传播对其进行优化。 具体来说，给定模板，T = {[T0:i ],[MASK],[Ti+1: j ]}，它不同于传统的离散提示，满足 [Ti ] ∈V 并将 T 映射为：

DART 将 [Ti] 视为伪标记并将模板映射如下：

其中 hi(0 ≤ i ≤ j) 是可训练的参数。 可微模板优化可以获得超出原始词汇表 V 的表达模板。 最后，模板 hi 通过以下方式进行了差异优化：

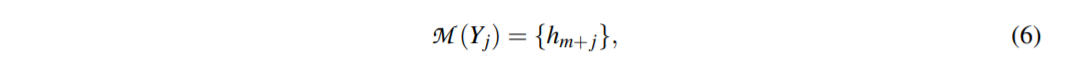
请注意，提示嵌入的值 hi 必须相互依赖而不是独立。 与使用双向 LSTM 的 P-tuning [26] 不同，DART 利用没有任何外部参数的辅助流畅性约束目标将提示嵌入相互关联，从而刺激模型专注于上下文表示学习。

## 4.3 Differentiable Label Optimization

基于提示的微调需要填写一个单词，并将被屏蔽的单词预测映射到一个语言器，生成一个类（即“是”：真。“否”：假）。对于每个类 c ∈ Y，之前的方法（例如 LM-BFF [14]）估计初始 L 在前 k 个词汇词的修剪集 V c ⊂V 上的条件似然。

然而，蛮力标签搜索：（1）计算量大且繁琐，因为 Ddev 通常非常大，需要多轮评估。 (2)随着类数的增加，可扩展性差（很多分类数据集有100多个类），搜索次数可能是kC（C代表类的总数），这是指数级的，因此难以处理。此外，类的标签包含丰富、复杂的语义知识，一个离散的标记可能不足以表示这些信息。

具体来说，对于标签 Y = {Y1,Y2,..,Ym}，不同于之前将类类型 Yi 转换为可变数量的标签标记 {...,v1,..,vk ,. ..}，DART 将 Yj 映射到一个连续的词汇空间，如下所示：



，其中 m 是模板中可训练嵌入的数量。 为避免优化任何外部参数，将 {h1,...,hm,..,hm+n} 替换为 V 中未使用的标记（例如，[unused1] 或词汇表中的特殊标记）以生成 V 0 ，如下所示 图1。

## 4.4 Training Objectives

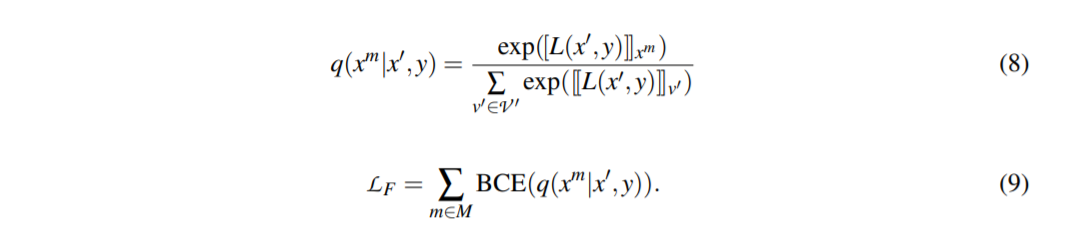
由于提示模板中的伪标记必须相互依赖，因此我们引入了辅助流畅性约束训练，而没有优化受 [26, 42] 启发的任何其他参数。 总体而言，有两个目标：类别区分目标 LC 和流畅性约束目标 LF。

类别判别目标类别判别目标是对句子进行分类的主要目标。 如图 1 所示，给定 (Xin,T)，我们可以将 Xprompt 生成为：

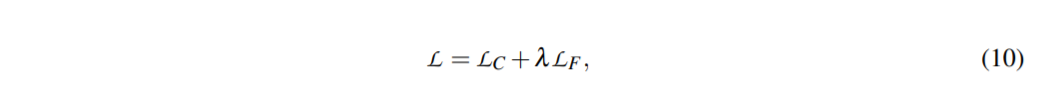


其中 CE 是交叉熵损失函数，LC 代表类别区分损失。

Fluency Constraint Object 为了确保模板标记之间的关联并保持从 PLM 继承的语言理解能力，我们利用 MLM 的流畅性约束对象。 如图 1 所示，输入句子中的一个 token 被随机屏蔽，并进行屏蔽语言预测。 x 和 x 0 分别是原始序列和掩码序列。 设 x m 为在 x 0 中被屏蔽的目标标记，P(x m|x 0 , y) 最大化如下4：



通过优化 OF，语言模型可以获得更好的上下文表示，模板标记之间具有丰富的关联。 我们有以下训练对象：



其中 λ 是超参数。 最后，我们介绍了 DART 的整体优化过程。 为了减轻少镜头微调的不稳定性，我们在算法 1 中引入了一个两阶段优化算法。 在优化步骤（第 2-6 行），{h0,...,hm,..,hm+ §4.2 和 §4.3 中描述的 n} 首先优化以获得最佳提示（我们通过停止其他参数的梯度来实现该过程），然后在步骤中优化所有参数（第 7-11 行）。

## 4.5 Comparison to Previous Prompt-tuning Approaches

由于即时学习已成为人与 PLM 交流的新范式或一种新方式，因此它吸引了许多研究人员。 由于快速学习的快速发展，不同的研究团队可能会引入一些类似的想法（学习嵌入）。 但是，我们列出了我们的模型与其他方法之间的主要区别，如表 1 所示：

总而言之，我们的方法非常简单，不需要外部参数（不同于 WARP/Prefix-Tuning/P-tuning/ADAPET）。 此外，我们的方法统一了模板和答案的优化。

# 5 Experiments

在本节中，我们详细介绍了在分类任务上进行的综合实验结果。 有希望的结果表明，我们提出的 DART 大大优于传统的微调方法，从而使预训练的语言模型成为更好的少样本学习者。

## 5.1 Dataset Statistics

我们对 15 个 NLP 任务进行了全面研究，涵盖了情感分析、自然语言推理、释义、句子相似性、关系提取和事件提取（我们只报告事件参数提取性能）。 评估包括 10 个流行的句子分类数据集（SST-2、MR、CR、Subj、TREC、MNLI、SNLI、QNLI、MRPC、QQP）。为了进一步评估所提出方法在复杂标签空间下的有效性，我们进行了实验 在关系提取和事件提取数据集上，包括 SemEval-2010 Task 8 [18]、TACRED-Revisit [1]、Wiki805 [16]、ChemProt [20] 和 ACE-20056。

## 5.2 Settings

所提出的模型是使用 Pytorch [31] 实现的，我们的实验是在 LM-BFF [14] 之后使用相同的设置进行的，该设置使用一组固定的种子 Sseed 测量每个任务的五个不同采样 Dtrain 的平均性能。 我们利用对多个超参数的网格搜索，并为每个集合 {D s train,Ddev},s ∈ Sseed 选择在 Ddev 上测量的最佳结果。 我们使用 AdamW 作为优化器。 我们使用 RoBERTa-large [27] 对分类任务进行实验，以便与 LM-BFF 进行公平比较。 除了我们对 ChemProt 数据集使用 SCIBERT [5] 之外，我们利用 uncased BERT-large [10] 来进行关系提取数据集。 我们遵循 [41] 并统一使用特殊的实体标记来突出显示用于关系提取的实体提及。

## 5.3 Main Results

如表 2 所示，我们观察到我们的方法获得了比传统微调更好的性能，并获得了与 LM-BFF 相当的结果。请注意，DART 不需要任何即时工程或外部模型（例如 LM-BFF 中的 T5）来生成易于适应其他数据集的模板。 DART 在 MR 数据集上每类仅 16 个训练样本即可获得 11.3% 的改进，与利用 T5 生成适当提示的 LM-BFF 相当。这些结果表明，DART 可以更好地激发潜在能力，并使预训练的语言模型成为更好的少样本学习器。我们还注意到 DART 比 P-tuning 产生更好的性能，这表明标签优化是有益的。

对于具有复杂标签空间的分类任务，如表 3 和图 2(a) 所示，我们观察到 DART 在关系提取和事件提取数据集上优于传统的微调方法以及 LM-BFF在少镜头和完全监督的设置中。所提出的方法在完全监督的情况下在 TACRED-Revisit 数据集上实现了 2.8% 的绝对性能改进，并且在每类只有 8 个训练样本的情况下产生了 18.4% 的收益。这些发现还表明，无需专家干预即可确定更相关的模板和标签，从而可以将所提出的方法推广到其他领域。此外，我们注意到当 K 变大（即从 8 到 32）时，改进会缓慢衰减。我们的方法是一种简单而有效的微调范式，不需要在复杂的标签空间内进行即时工程，因此，可以成为某些 SOTA 模型的合适插件。

## 5.4 Ablation Study

我们进行消融研究以验证所提出方法中组件的有效性。 我们观察到，在缺少任何一个模块（即流畅性约束对象、可微模板或可微标签）的情况下，DART 表现出性能下降，这表明所有模块都是有利的。 此外，我们注意到可微标签优化对性能更敏感，并且对 DART 非常有益，尤其是对于低资源设置。 由于所提出的方法是第一种利用可微标签优化的方法，因此这些发现说明合适的标签标记很重要。

## 5.5 Analysis and Discussion

DART 可以应用于其他预训练的 LM 吗？

为了评估所提出的方法是否可以应用于其他 LM，我们使用 GPT-2-medium 进行了实验。从图 2(b) 中，我们观察到使用 GPT-2 介质的 DART 比传统的微调方法产生更好的性能。此外，我们注意到使用 GPT-2-medium 的 DART 可以达到与 BERT-large 相当的性能，如 [26] 所观察到的那样，表明 GPT 式架构在自然语言理解方面的潜力被低估了。

什么是完全优化的提示？

由于所提出的方法中的提示模板和标签标记被映射为 {h1,...,hm,..,hm+n}，我们进一步分析了究竟是什么优化了标签学习。我们进行最近邻词汇嵌入搜索，将 V 中的 Top-3 优化伪标签标记投射到可读的自然语言中。我们使用 t-SNE [43] 和标准化来可视化 Wiki80 数据集上的标签。例如，“military\_branch”指的是 red ?图 3 中的关系类型是通过优化连续空间中的伪标签来学习的，“志愿者”、“下士”和“伙伴”，指的是 • 是最接近标签的标记。这一发现表明优化的标签嵌入可以呈现更好的语义表示能力。

DART 与常规微调

所提出的方法执行小样本学习的能力可以归因于标签并且是一个真正的语言理解任务，一旦模型能够正确执行它，它就可以轻松地将这些知识应用到其他任务中这样的。从表面上看，(i) DART 没有优化任何新参数；然而，传统的微调应该在 [CLS] 嵌入上学习一个明确的分类器，这在低数据情况下可能会失败。 (ii) DART 具有与大规模语言模型预训练相同的任务设置，并且对于下游分类任务具有较小的理论上限 [36]。

限制

当任务语料库的分布与预训练语料库的分布不同时，我们的工作可能会失败。例如，通用的预训练语言模型可以在特定领域（例如，医学领域）中使用更多训练实例进行微调。这个问题可以通过中级培训来解决 [33, 46, 51]，并将在未来的工作中进行分析。此外，我们的工作还显示了与超参数相关的不稳定性，[12, 50, 32] 也观察到了这一点，作为 NLP 中少样本学习的波动性。然而，总的来说，我们相信我们的工作将激发未来的工作，将更多的实际应用应用于低数据设置，例如，涉及低资源语言或专家注释。

# 6 Conclusion and Future Work

本文介绍了 DART，这是一种简单而有效的微调方法，可改进快速学习预训练语言模型。与传统的微调方法相比，所提出的方法可以在少拍场景中产生令人满意的改进。所提出的方法也可用于其他语言模型，并可扩展到其他任务，例如意图检测。直观地说，本研究中获得的结果可用于激发 NLP 小样本学习的两个未来研究方向：（i）将所提出的方法扩展到半监督设置以进一步利用未标记数据； (ii) 将所提出的方法扩展到少拍终身学习，而提示必须通过自适应任务进行优化。

更广泛的影响

预训练微调方法已成为自然语言处理 (NLP) 的标准。然而，监督微调实际上仍然受到标记数据的影响。本研究提出了一种名为 DiferntiAble pRompT (DART) 的新型可插拔、可扩展和高效方法，该方法可以将小型语言模型转换为更好的少样本学习器，而无需任何提示工程。我们相信我们的研究对文献做出了重大贡献，因为确定适当的提示需要领域专业知识，而手工制作高性能提示通常需要不切实际的大型验证集，并且使用所提出的方法已经克服了这些问题，与模型无关、参数高效且独立于即时工​​程。我们在 13 个标准 NLP 任务上实验验证了我们提出的方法，并且它被认为优于几个标准 NLP 平台。