ABSTRACT

“指令微调”可以增强LLMs的能力。这种方法通常需要从大量数据中找到最相关的部分（指令数据），以便培养模型的特定技能，比如推理能力。

论文提出LESS算法，用于估计数据影响并执行低秩梯度相似性搜索以选择指令数据。更重要的是，LESS兼容现有的影响力公式，以与Adam 优化器和可变长度指令数据一起使用。

LESS 首先利用低维梯度特征构建了一个高度可复用和可迁移的梯度数据存储库，然后选择那些与“体现特定能力的少量样本“相似的样本。

实验表明，在不同下游任务中，对选择的 5% 数据进行训练通常可以优于对完整数据集进行训练。此外，所选数据具有高度可转移性：即可以利用较小的模型为较大的模型和来自不同系列的模型选择有用的数据。定性分析表明，作者方法超越了表面形式线索，可以识别能够体现预期下游应用所需推理技能的数据。

1 INTRODUCTION

指令微调使LLMs能够熟练遵循人类指令。近期研究都主要致力于增加指令调整数据集的多样化和广泛性，使模型从少量样本也能获得较强的泛化能力。但使用这种混合指令数据集训练LLMs会阻碍其特定能力的发展。例如，在混合指令微调数据集上训练的 LLMs 表现出的性能比在数据子集上训练的性能更差。

此外，考虑到广泛的用户查询以及响应这些查询所需的多种技能，可能并不总是有足够的域内数据可用。因此，作者希望能够**有效地利用通用指令微调数据来提高特定能力。**

因此作者提出问题：

> 仅给出**少数体现特定能力的样本**，我们如何从大量指令数据集中有效地选择相关的微调数据？

作者提出有针对性的指令调整来解决这一问题。方法是，优先使用在目标任务中使得损失最小的数据进行训练，而不是依赖surface form features (表面形式特征，比如词长、词频、词序、标点符号使用等，Gururangan et al., 2020; Xie et al., 2023b)。

受到过去的工作**利用梯度信息估计单个训练数据点影响**的启发，作者设计了优化器感知方法来选择数据。但是这种影响力公式的直接应用面临着指令调优设置特有的几个挑战：

- LLMs传统上是使用 Adam 优化器而不是过去工作中考虑的标准的 SGD 优化器进行微调；
- 使用可变长度指令数据的序列级梯度可能会导致破坏影响估计；
- LLMs中大量的可训练参数使得梯度的计算和存储极其消耗资源。

作者在 LESS 中解决了这些问题，LESS 是一种执行低秩梯度相似性搜索来为目标应用程序选择相关指令微调数据的算法，它具有以下属性：

1. 与 Adam 的指令微调兼容：LESS 采用经典影响力公式中的梯度特征，以与 Adam 优化器和可变长度指令数据配合使用。优化洞察力和影响力公式也可能具有独立的意义。
2. 高效：LESS 使用 LoRA 和随机投影构建具有低维、易于操作的梯度特征的梯度数据存储，从而允许高效且有效的数据集选择。梯度数据存储可以重复用于新的目标任务。
3. 可转移：使用小模型的梯度特征选择的数据会在大型模型和来自不同系列的模型中产生强大的性能，从而提高 LESS 的效率。
4. 可解释：定性分析表明，LESS 选择具有相似推理和技能类型的数据作为目标任务，而现有方法通常根据表面形式线索（例如：语言或主题）选择数据。

作者在三个不同的下游数据集（MMLU、TYDIQA 和 BBH）上评估我们的方法，每个数据集都包含不同的子任务，可以有效地模拟目标指令微调场景。

结果：

LESS 通常会选择一小部分数据子集 (5%)，其性能优于整个数据集上的训练，并且所选子集在不同模型参数规模和模型系列中仍然普遍有效。

与其他数据选择方法的比较表明，**LESS 是唯一一致有效的方法**，证明其相对**较高的计算成本**是合理的。

2 ESTIMATING THE INFLUENCE OF INSTRUCTIONS

数学部分，我是菜狗

大概东西：
使用训练动态的一阶近似（泰勒展开）来**估计训练数据点对保留数据的影响。**

考虑了每一步的影响（模型在时间步上的训练损失）、轨迹影响

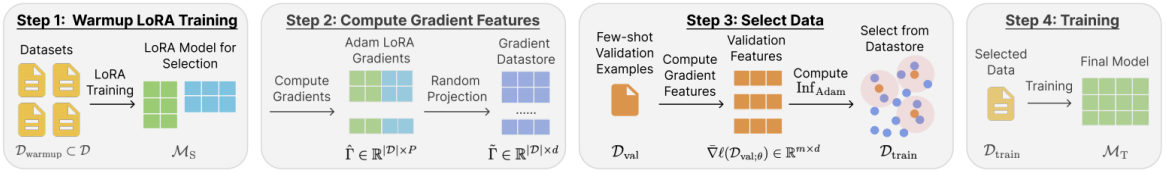
影响力公式：

$$\text{Inf}_{\text{SGD}}(z, z') \triangleq \sum_{i=1}^{N} \bar{\eta}_i \langle \nabla \ell(z'; \boldsymbol{\theta}_i), \nabla \ell(z; \boldsymbol{\theta}_i) \rangle \qquad (1)$$

4 LESS: ESTIMATING INFLUENCES EFFICIENTLY

第三节确定了如何**使用模型梯度来估计影响**，但考虑到 LLMs 的庞大规模，计算和存储模型梯度仍然非常昂贵。

LESS计算步骤：



第一步：**热身阶段。**输入数据集的一个子集，用LoRA对预训练的基础模型（例如LLAMA-2-7B）进行参数高效的微调，得到一个选择模型M，这步可以减少可训练参数的数量并加速训练过程。

在训练集的一个随机子集上进行N个epoch的预热训练，以适应特定的数据分布，并在每个epoch后保存模型检查点。

第二步：**计算梯度特征。**为每个候选数据点计算预热训练期间的梯度，然后将这些梯度投影到低维空间，生成低维梯度特征，保存在梯度数据存储库中，以便后续的数据选择过程可以高效地重用这些特征。

第三步：**数据选择。**对于目标任务的验证集（包含少量样本），计算每个子任务的梯度特征，使用LESS算法计算每个训练数据点对于验证集的潜在影响，通过评估数据点的梯度特征与验证集特征之间的相似性来打分。

根据得分选择最高的一部分训练数据点（例如前5%）作为最终的训练集。

第四步：**目标模型训练。**在目标模型M上使用所选数据D _{train}进行最终训练，该模型可以使用可以使用LoRA或完全微调进行训练。

训练完成后，使用目标模型在测试集上进行评估，以验证LESS算法选择的数据对于提升模型性能的效果。

整个过程的核心思想是利用模型的梯度信息来估计数据点对于目标任务的影响，并通过选择具有高影响力的数据点来进行有针对性的训练，从而提高模型在特定任务上的性能。LESS方法的关键在于它能够适应现有的优化器（如Adam），并且能够有效地处理可变长度的指令数据。此外，LESS构建的梯度数据存储库可以重用于不同的目标任务，提高了数据选择过程的效率。

验证集：

验证集是固定的，每个子任务只包含几个样本，且验证集是源数据集自带的。

5 实验

训练数据集，使用以下指令调整数据集：

(1) 从现有数据集创建的数据集，例如：FLAN V2 (Longpre et al., 2023) and COT (Wei et al., 2022c)；

(2) 具有人工编写答案的开放式生成数据集，包括 DOLLY (Conover et al., 2023) 和 OPEN ASSISTANT。

这些数据集的格式和底层推理任务差异很大。训练数据集不包含目标查询的任何明显的域内数据。

评估数据集。我们在 MMLU (Hendrycks et al., 2020), TYDIQA (Clark et al.,2020) 和 BBH (bench authors, 2023) 上评估我们的方法。

MMLU 包含多项选择题，涵盖 57 个任务，包括初等数学、美国历史、计算机科学、法律等。

TYDIQA 是一个多语言问答数据集，包含 11 种类型不同的语言。给定一个问题和相关段落，该任务需要从段落中提取答案。

BBH 是 BIG-Bench 中用于评估推理能力的 27 项挑战性任务的集合。

下表包含有关这些任务的更多详细信息。每个数据集都包含多个子任务，每个子任务都带有少量示例。这些示例用作数据选择（第 4.2 节）的 DvalD_{val}Dval 以及评估中的少量上下文学习演示。

Table 1: Statistics of evaluation datasets. The selection of evaluation tasks cover different kinds of answer types.

Dataset	# Shot	# Tasks	# Instances	Answer Type
MMLU	5	57	18,721	Letter options
TYDIQA	1	9	1,713	Span
BBH	3	23	920	COT and answer

用于数据选择和训练的模型：

我们测试LESS具有三个base模型：LLAMA-2-7B ， LLAMA-2-13B 和 MISTRAL-7B。在迁移设置（如：LESS-T）中，我们数据选择使用LLAMA-2-7B作为MsM_sMs，并训练LLAMA- 2-13B 或 MISTRAL-7B 作为目标模型 MTM_TMT 。预热训练和最终模型训练均使用LoRA进行。我们的报告中使用了三个随机种子的平均性能和标准差。

默认设置：

LESS 对随机选择的完整数据集的 5% 进行 4 个 epoch 的预热训练，并计算数据 D 上的 8192 维梯度特征（第 4.1 节）。对于每个目标任务，我们使用这些特征根据其影响对数据点进行评分（定义 3.1）并选择DDD中得分最高的5%来构建DtrainD*{train}Dtrain。我们在这个选定的数据DtrainD*{train}Dtrain上训练目标模型MTM_TMT 。

关键结论：

- LESS在不同的模型中都是有效的，在所有模型和评估数据集中，**LESS筛选的数据始终显著优于随机**；
- 选择的**5%高价值数据通常优于整个数据集的效果**
- **使用小模型选择的数据可以提高较大和不同模型的性能**
- 与其他 baseline 相比 LESS 是唯一持续有效的方法

总结

本文提出了一种基于优化器感知影响力的数据选择算法LESS。 LESS 创建了一个有效且可重用的低维梯度特征的数据存储，以实现高效的数据选择。实验证明了 LESS 与全量数据(100%)、随机数据(5%)相比的有效性，并强调了**使用较小模型选择数据来训练较大模型的能力。**分析和消融实验表明，本文的方法**选择了更多可解释的数据，但计算成本可能很高。**