计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(1):276−284 [doi: 10.15888/j.cnki.csa.009731] [CSTR: 32024.14.csa.009731]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于奇异值分解的适应微调①

林志鹏, 郭峥嵘, 张伟志, 郭躬德

(福建师范大学 计算机与网络空间安全学院, 福州 350109)
通信作者: 郭躬德, E-mail: ggd@fjnu.edu.cn

摘　要: 大语言模型的兴起对自然语言处理领域产生了深远影响. 随着计算资源的增长和模型规模的扩大, 大语言模型在自然语言处理中的应用潜力日益显现. 然而, 广泛使用的低秩适应微调方法在面对模型规模增大时, 遇到了微调效率和存储成本等方面的挑战. 为了解决这一问题, 本文提出了一种基于奇异值分解的适应微调方法. 该方法只需将奇异值分解得到的对角矩阵和缩放向量作为可训练参数, 从而在降低训练成本的同时, 实现了在多个自然语言处理任务上的性能提升. 实验结果显示, 基于奇异值分解的适应微调方法在 GLUE 和 E2E 基准测试中的性能超越了同等数量级的方法. 通过与常用的参数高效微调方法进行比较, 发现基于奇异值分解的适应微调方法在减少可训练参数数量和提高微调效率方面具有显著优势, 并在可训练参数微调效率实验中实现了最高的性能增益. 在未来的研究中, 将专注于进一步优化基于奇异值分解的适应微调方法, 在更广泛的任务和更大规模的模型中实现更高效的微调.

关键词: 参数高效微调; 生成式大模型; 深度学习; 领域适配; 有限算力

引用格式: 林志鹏,郭峥嵘,张伟志,郭躬德.基于奇异值分解的适应微调.计算机系统应用,2025,34(1):276–284. http://www.c-s-a.org.cn/1003-3254/9731.html

Adaptation Fine-tuning Based on Singular Value Decomposition

LIN Zhi-Peng, GUO Zheng-Rong, ZHANG Wei-Zhi, GUO Gong-De

(College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350109, China)

Abstract: The rise of large language models has profoundly impacted natural language processing. With the growth of computational resources and the expansion of model sizes, the potential applications of large language models in natural language processing are increasingly evident. However, the widely used low-rank adaptation (LoRA) method faces challenges related to fine-tuning efficiency and storage costs as model sizes increase. To address this issue, this study proposes a singular value decomposition-based adaptation fine-tuning method. This method only requires the diagonal matrix and scaling vector obtained from singular value decomposition to be trainable parameters, achieving performance improvement in multiple natural language processing tasks while reducing training costs. Experimental results show that the proposed method outperforms other methods of the same order of magnitude in GLUE and E2E benchmark tests. Compared with commonly used parameter-efficient fine-tuning methods, it demonstrates significant advantages in reducing the number of trainable parameters and improving fine-tuning efficiency, achieving the highest performance gains in experiments on the fine-tuning efficiency of trainable parameters. Future research will focus on optimizing the proposed method to achieve more efficient fine-tuning in a wider range of tasks and larger-scale models.

Key words: parameter efficient fine-tuning (PEFT); large generative model; deep learning; domain adaptation; limited computational resource

① 基金项目: 国家自然科学基金 (61976053, 62171131)
收稿时间: 2024-06-06; 修改时间: 2024-07-10; 采用时间: 2024-07-18; csa 在线出版时间: 2024-11-25
CNKI 网络首发时间: 2024-11-26

近年来, 生成式模型在多个领域内取得了重要成果. 学术界和工业界对 LLM 的研究都得到了很大程度的推进, 其中令人瞩目的进展是 ChatGPT 的推出, 引起了社会的广泛关注. GPT-3、ChatGPT、LLaMA、文心一言、通义千问等模型, 借助亿级的参数量与海量的训练数据, 在许多领域任务上得到了惊艳的效果[1]. 目前 GPT-4 与 PaLM 2 在医学 AI 创新的前沿占据了中心地位, 这些新技术在临床、教育和研究工作中得到了广泛的应用[2].

预训练微调策略通过在通用预训练模型上进行少量调整, 大幅节省了计算资源, 并在很大程度上促进了大型模型的成功. 传统的微调使用预先训练的模型来生成上下文表示, 然后添加特定任务的分类层. 这种在下游数据集上使用全部参数微调的方法可以显著提升模型性能. 然而, 随着模型参数规模的持续扩大, 全参数微调的局限性逐渐显现. 特别是在计算资源和存储需求方面, 这对于大规模部署和实时应用带来了巨大的挑战. 此外, 针对下游任务的全参数微调训练过程中, 使用大量的参数会导致过拟合和泛化能力下降等问题. 这也极大地限制了大模型在各种垂直场景中的高效灵活适配.

参数高效微调 (parameter efficient fine-tuning, PEFT) 技术旨在保持模型泛化能力的同时, 仅通过微调少量参数来适应特定任务. PEFT 技术的关键在于冻结原模型的大部分参数, 仅微调小部分参数或引入少量额外的训练参数, 从而减少过拟合的风险[3]. 与传统的全参数微调相比, PEFT 更加高效, 适用于大规模模型的部署和实际应用.

然而, Ding 等人[4]研究了 1 200 篇已发表的 NLP 领域论文, 发现大约 0.5%–4% 的研究采用了超过 10 亿参数的预训练模型. 这些模型仍不受欢迎的一个主要原因是其部署和实验验证成本难以承受. 作为使用最为广泛的 PEFT 方法, LoRA[5]在面对这些模型时, 内存限制成为一个显著的障碍. 这样庞大的内存需求不仅在计算资源上构成了一个巨大的挑战, 同时也引发了存储成本的激增.

为了解决这一难题, QLoRA[6]在 LoRA 的基础上使用了 4 bit 量化技术, 进一步减少了微调过程中的显存占用. 尽管这种方法有效地避免了在梯度检查点期间由于内存峰值导致的内存不足问题. 但 QLoRA 通过牺牲少量精度来减少存储需求, 并没有实现对可训练参数的高效利用. 这意味着尽管内存使用得到了优化, 但模型性能可能会因训练时精度的降低而受到影响, 未能充分发挥大型模型的潜力.

综上所述, LoRA 在传统方法上通过减少参数数量和内存需求, 实现了更高效的微调. 然而, QLoRA 尽管在内存优化方面取得了显著进展, 但在可训练参数的高效利用上仍存在不足. 因此本文提出了一种基于奇异值分解的适应微调方法 (singular value decomposition based adaptation, SvdA), 通过使用缩放向量提高了微调效率. 本文主要贡献如下.

1) 提出了一种基于奇异值分解的参数高效微调方法 SvdA, 无需额外的推理时间成本, 并减少了 LoRA 等传统方法的可训练参数的数量.

2) 为了验证 SvdA 不使用秩的策略是否能带来更高的实验效率, 与基于向量的随机矩阵适应方法 (vector-based random matrix adaptation, VeRA) 进行了对比实验. 实验结果显示, SvdA 方法在无需定义秩的前提下, 能够获得更优异的实验结果. 这表明 SvdA 方法在简化实验设置的同时, 仍能取得更好的性能.

3) 将 SvdA 方法与 LoRA、VeRA 等微调方法在自然语言理解 (GLUE) 和自然语言生成 (E2E) 基准上进行了实验对比, 证明了 SvdA 方法的有效性.

4) 研究各参数高效微调方法在 RTE 任务上的性能增益效率. 实验结果表明, SvdA 每增加 1k 的可训练参数, 其获得的性能增益最高.

1 相关工作

大型预训练模型的出现不仅提升了模型性能, 同时也使得微调工作变得更具挑战性. 在当前资源受限的环境中, 维护和更新这些模型变得尤为困难. 为了应对这一挑战, 参数高效微调方法成为一种有效的策略, 将大型预训练模型适配到不同的下游任务[7]. 仅通过优化少量参数, 实现高效地驱动大型语言模型的微调任务, 从而降低由于参数规模不断扩大所带来的微调和存储成本.

具体而言, 参数高效微调方法包括 4 种类型[8]: 选择性方法通过选择性地更新预训练模型的一部分参数来实现微调; 适配器加性方法引入可以在预训练模型不同层中添加的适配器层, 以便在微调时进行参数更新; 软提示词加性方法[9,10]使用软提示词来指导微调, 用以调整模型的输出; 而重新参数化方法[11]通过重新

定义模型的参数来实现微调, 其中一种方式是将权重分解为多个部分, 以便灵活地适应微调任务. 如图 1 所示.

重新参数化方法是目前被广泛采用的参数高效微调方法, 相对于其他类型的方法, 它使用较少的可训练参数量, 取得了传统微调相当的性能. Aghajanyan 等人[12]提出了一种名为 Intrinsic SAID 的重新参数化方法, 该方法通过 Fastfood 变换将高维向量投影到低维空间, 从而重新参数化模型权重的更新. 他们的研究发现, 常见的预训练模型存在一种低维度的重新参数化方式, 使得微调效果与全参数微调相当. 然而, 由于该方法需要对所有模型参数进行更新, 因此在微调大型网络时并不适用.

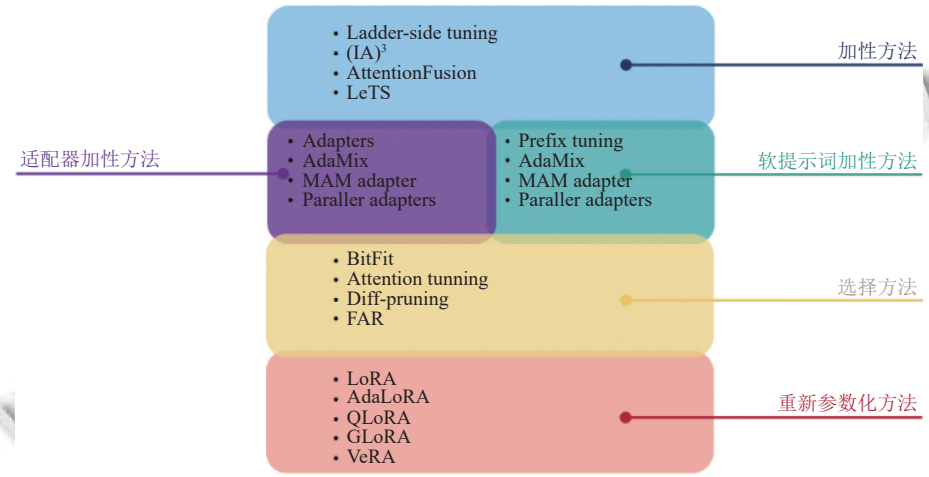

图 1 参数高效微调方法汇总

低秩适应微调 (LoRA) 的核心思想是将预训练模型的权重冻结, 并将可训练的低秩分解矩阵注入每层 Transformer 架构中. 具体来说, 它在原始权重 $W^{(0)}\in R^{d1\times d2}$ 旁路增加两个可训练的低秩矩阵, 第 1 个矩阵 $A\in R^{r\times d2}$ 负责降维, 第 2 个矩阵 $B\in R^{d1\times r}$ 负责升维, 中间层维度 r 用来模拟微调后矩阵的本征秩, 且 $r\ll\min(d1, d2)$. 其中 $d1$、$d2$ 表示原模型层的输出维度和输入维度, 式 (1) 中的 x、h 与 $b^{(0)}$ 表示输入、输出特征和偏置, $\Delta W\in R^{d1\times d2}$ 表示该层的增量更新矩阵. 对于 $h=W^{(0)}x+b^{(0)}$, 修改后的正向传递为:

$$\begin{aligned} h &= W^{(0)}x+\Delta Wx+b^{(0)}\\ &= W^{(0)}x+BAx+b^{(0)} \end{aligned} \tag{1}$$

在一项研究中[13], LoRA 方法使用 2M 指令数据对 LLaMA-7b 模型进行微调, 采用 8 个 NVIDIA A100-40 GB GPU. 结果显示, LoRA 方法完成一轮训练需 7 h, 而全参数微调需 31 h. LoRA 显著提升了时间效率, 但随着模型规模的增长, 寻找节省存储和算力的微调方法变得愈发重要.

自适应预算分配的低秩适应微调 (AdaLoRA)[14]的目标是根据每个参数的重要程度自动分配可微调参数的预算. AdaLoRA 也尝试采用奇异值分解的形式来参数化增量矩阵. 这种参数化方式不仅规避了大量 SVD 运算, 还允许高效地裁剪不重要的更新中的奇异值, 从而降低计算资源的消耗. 式 (2) 中的 $P\in R^{d1\times r}$、$Q\in R^{r\times d2}$ 和对角矩阵 $\Lambda\in R^{r\times r}$ 分别为奇异值分解得到的 3 个矩阵, Λ 中包含奇异值 $\{\lambda_i\}_{1\leqslant i\leqslant r}$ 且 $r\ll\min(d1, d2)$, 其他符号的含义与 LoRA 中类似. 为了确保在训练开始时 Λ 被初始化为 0, P 和 Q 采用随机高斯初始化. 对于 $h=W^{(0)}x+b^{(0)}$, 修改后的正向传递为:

$$\begin{aligned} h &= W^{(0)}x+\Delta Wx+b^{(0)}\\ &= W^{(0)}x+P\Lambda Qx+b^{(0)} \end{aligned} \tag{2}$$

AdaLoRA 不仅拥有比 LoRA 更少的可训练参数, 而且在多个指标上优于 LoRA. 奇异值分解方法在其中发挥着重要作用. 此前的研究表明, 奇异值分解方法在神经网络的各个领域中起着重要作用, 特别是在神经网络的优化和训练中展现了其巨大的潜力[15–20]. 虽然 AdaLoRA 考虑了节省奇异值分解中的计算开销, 但相比 LoRA, 其可训练参数量并没有显著减少. 类似的方法还有 PiSSA[21], 它仅在参数初始化过程中使用了奇异值分解, 但本质上并未改变 LoRA 的结构. 因此, 上述方法在微调任务中仍面临存储和算力不足的困境.

基于向量的随机矩阵适应 (VeRA)[22]并没有选择和先前研究一样将旁路矩阵全部作为可训练参数, 而是将 LoRA 中的增量矩阵全部冻结, 并插入两个的缩放向量来适应低秩矩阵. 在适应下游任务的过程中, VeRA 仅使用缩放向量作为可训练参数. 在训练过程中, 共享矩阵 AB 初始化后保持静态. 共享矩阵并不会占用额外的存储空间, 因此 VeRA 中的 r 并不需要是低秩的. 下面公式中 $\Lambda_b \in R^{d1\times d1}$, $\Lambda_d \in R^{r\times r}$, 其他符号的含义与 LoRA 中类似. 对于 $h = W^{(0)}x + b^{(0)}$, 修改后的正向传递为:

$$\begin{aligned} h &= W^{(0)}x + \Delta W x + b^{(0)} \\ &= W^{(0)}x + \Lambda_b B \Lambda_d A x + b^{(0)} \end{aligned} \tag{3}$$

上述的重新参数化的参数高效微调方法都采用了可训练矩阵与冻结权重可合并的结构, 因此它们在部署时不会产生额外的推理时间成本. 当前, 参数高效微调领域的研究重点仍集中在如何有效地减少可训练参数. 为应对这一挑战, VeRA 引入了缩放向量的概念, 从而显著降低了实验研究的成本. Zhang 等人[14]的研究中提到, LoRA 或 VeRA 中的 A, B 矩阵并非正交矩阵, 这意味着其行或列之间的特征具有较高的相关性. 当秩降低或是特征置零时, 这种特征相关性会被破坏. 奇异值分解方法因其优越的数学性质在训练中表现出更强的鲁棒性, 但庞大的计算量增加了微调任务的负担. 因此, 需要进一步研究如何在参数高效微调中合理应用奇异值分解方法.

2 基于奇异值分解的适应微调

在参数高效微调领域, 有效减少可训练参数一直是研究的重点. 现有方法通常采用可训练矩阵与冻结权重可合并的结构来实现参数高效微调, 但仍存在一些挑战. 本文提出基于奇异值分解的适应微调 (SvdA) 方法, 该方法使用了缩放向量和奇异值分解对角矩阵这两个一维的向量作为可训练参数, 有效地平衡了存储成本与性能之间的关系, 为将奇异值分解有效应用于参数高效微调领域开辟了新的途径.

2.1 基本定义

与先前方法相同, SvdA 冻结了原始权重矩阵 W, 但它并没有采用先前的低秩矩阵作为可训练参数或是共享矩阵的策略. 具体而言, SvdA 冻结了通过奇异值分解生成的两个奇异向量矩阵, 并将对角奇异值矩阵 Λ_s 改为缩放向量的形式. 随后, 在奇异值分解矩阵之前插入了一个新的缩放向量 b. 该方法包含两个重要组成部分.

1) 奇异值分解矩阵: SvdA 以奇异值分解原权重的形式构建增量矩阵, 使左奇异向量矩阵 U 与右奇异向量矩阵 V 作为缩放向量的共享矩阵. 通过奇异值分解保留了原权重矩阵的重要信息, 同时可以避免数值计算中的不稳定性问题. 在训练过程中, 这种方法能够保持较高的数值精度, 从而提高模型的稳定性和可靠性.

2) 缩放向量 (scaling vector): 缩放向量是用来表示对角矩阵的一维向量. 与 VeRA 中将缩放向量全部初始化为 1 的策略不同, SvdA 的缩放向量 s 是由奇异值分解得来的可靠数值. 为了增加微调模型的鲁棒性, SvdA 引入了第 2 个缩放向量 b, 通过两个缩放向量适应两个共享奇异向量矩阵, 详见图 2. 经过训练的缩放向量, 可以同共享奇异值向量矩阵合并到原始权重中, 从而消除额外的推理延迟.

SvdA 在本质上属于重新参数化的参数高效微调方法, 通过奇异值分解的共享矩阵与缩放向量作为原权重矩阵的旁路, 实现对模型的增量更新. 在生产环境中部署时, 可以显式地计算和存储 , 采用权重分解的策略在推理阶段不会带来额外的计算负担. 当 $d1>d2$ 时, $U\in R^{d1\times d2}$、$V\in R^{d2\times d2}$, 缩放向量 $s\in R^{1\times d2}$、$b\in R^{1\times d1}$. 当 $d1\leqslant d2$ 时, $U\in R^{d1\times d1}$、$V\in R^{d2\times d2}$, 缩放向量 $s\in R^{1\times d1}$、$b\in R^{1\times d2}$. 对于 $h = W^{(0)}x + b^{(0)}$, 修改后的正向传递为:

$$\begin{aligned} h &= W^{(0)}x + \Delta W x + b^{(0)} \\ &= W^{(0)}x + \Lambda_b U \Lambda_s V^{\mathrm{T}} x + b^{(0)} \end{aligned} \tag{4}$$

其中, 对角矩阵 Λ_s=diag(s)、Λ_b=diag(b), 其他符号的含义与 LoRA 中类似. 矩阵 U 和 V 是冻结且跨层共享的酉矩阵, 在训练过程中保持静态, 无需额外占据存储空间. 缩放向量 b 和 s 是可训练的, 通过学习这两个向量的值即可适应下游任务. SvdA 并没有直接将奇异值分解后的原权重矩阵作为可训练参数, 而是利用奇异值分解后得到的正交矩阵来保证微调任务的鲁棒性. SvdA 通过奇异值分解提取的奇异值来初始化缩放向量, 这不仅保留了原权重中的关键特征, 还显著加快了收敛速度, 并提升了训练稳定性. 这样的设计充分利用了奇异值分解在参数高效微调方法中的优势, 同时避免了额外的计算开销.

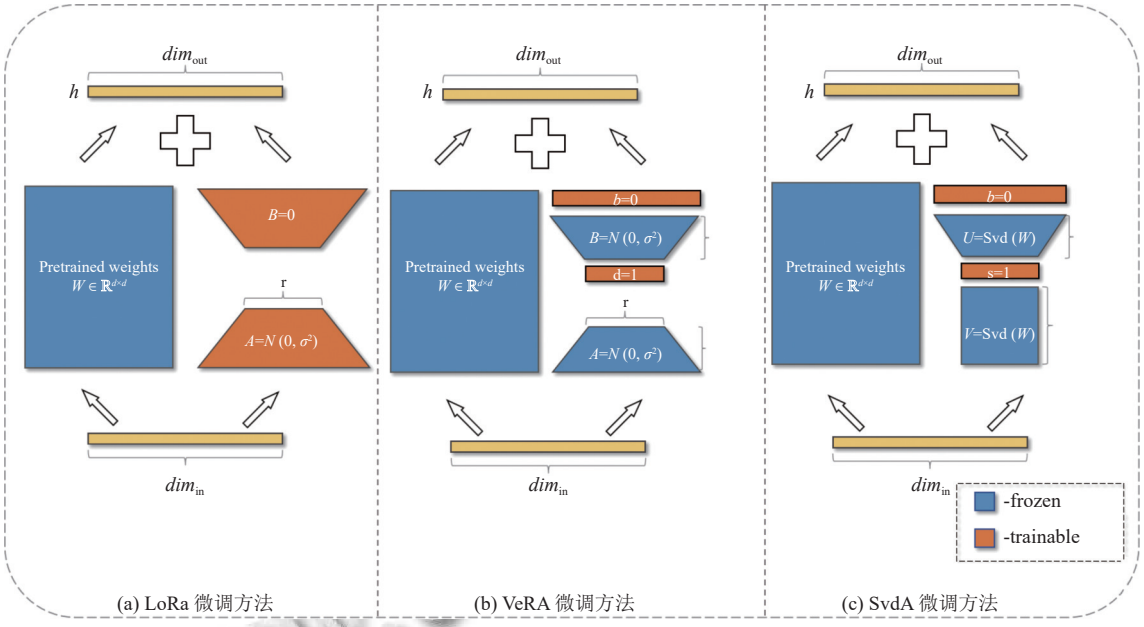

图 2 参数高效微调方法汇总

2.2 初始化策略

正交矩阵与矩阵相乘时会改变矩阵范数, 这一性质在解决梯度消失与梯度爆炸问题上具有重要意义[23]. Saxe 等人[24]的研究发现, 通过将权重矩阵初始化为正交矩阵, 深度线性神经网络可以更快地收敛到最优解. 在一些结合奇异值分解方法的神经网络的研究中[25,26], 通常采用正交约束来确保训练参数的正交性, 以更好地适应下游任务.

具体地, 在初始化过程中, SvdA 首先对原权重矩阵进行奇异值分解, 得到满足正交矩阵的特性的 U、V 矩阵. 与传统方法不同的是, SvdA 不会再对 U、V 矩阵进行额外更新约束, 而是将它们冻结, 从而确保共享矩阵的正交性. 这种初始化策略在保持模型有效性的同时, 提供了更好的训练稳定性和收敛速度.

为了确保权重矩阵在前向传播时不受影响, 通常将缩放向量 b 初始化为 0. 在模型微调任务中, 通过奇异值分解得到的对角奇异值矩阵 Λ_s 会被调整为缩放向量, 以适应微调过程中权重参数分解的需求. 缩放向量 s 在初始化时, 可以选择保留奇异值, 或将所有值置为非零值. 图 2 提供了 SvdA 中缩放向量初始化的示例.

2.3 数据集

E2E (end-to-end NLG challenge dataset)[27]作为一个在自然语言生成研究中被广泛采用的基准数据集, 其主要目标是推动端到端自然语言生成系统的进步. 该数据集包含大约 50 000 对的语义表示和对应的目标文本. 在这些对应关系中, 语义表示描绘了餐馆的各种属性, 而目标文本则是根据这些属性生成的自然语言描述.

在使用 E2E 数据集测试自然语言生成能力时, 通常会采用多种不同的评估指标. BLEU 指标[28]偏重于较短的翻译结果, 它更看重准确率而非召回率. NIST[29]是对 BLEU 的改进, 考虑了 n-gram 的罕见度, 使得罕见的 n-gram 匹配获得更高的权重. METEOR[30]通过考虑词形变化、同义词匹配以及词序来衡量生成文本与参考文本之间的相似度. ROUGE-L[31]则是一种基于召回率的相似性度量, 主要用于评估参考译文的充分性和忠实性. 而 CIDEr[32]则通 TF-IDF 加权的 n-gram 匹配来评估生成文本与参考文本之间的相似度.

GLUE 基准[33]由一系列自然语言理解数据集或任务组成. GLUE 的核心任务包括: 确定单句语法准确性的任务 CoLA; 确定单句情感的任务 SST-2; 检测两个句子是否是彼此释义的 MRPC; 确定两个问题是否语义等价的 QQP; 在连续尺度上测量两个句子的相似性的 STS-B; 在多样式环境中测试模型预测文本蕴含能力的 MNLI; 将斯坦福问答数据集转换为自然语言推断任务的 QNLI; 以及带有来自各种来源的数据的文本蕴含任务 RTE.

GLUE 和 E2E 基准在各自领域中推动了预训练模型的进步, 并通过不同的任务和指标为模型的全面测评提供了坚实基础. 这些基准不仅促进了模型在自然语言理解和生成方面的能力提升, 还为进一步研究和改进预训练模型的微调提供了宝贵的资源和参考框架.

3 实验

本节将进行一系列实验, 以评估 SvdA 方法的性能. SvdA 将与其他常见的微调方法进行比较, 测试它们在自然语言生成和理解基准数据集上的性能. 为了测试每增加 1k 参数对实验效果的提升率, 实验设置在 RoBERTa-base 模型[34]上对比不同参数高效微调方法在 GLUE 基准的 RTK 任务上的参数效率. 在 GLUE 基准中的 MNLI 任务上设置对比实验, 展示 SvdA 对比同量级方法无需寻找最优超参数秩的效率优势.

LoRA 方法通过采用缩放因子进行缩放, 能够在面对不同的 r 值时, 保持输出的大小一致. 当 r 值发生变化时, 这种策略有助于降低重新调整学习率的成本. 然而, 尽管引入缩放因子可以帮助维持稳定的输出大小, 但也可能增加过拟合的风险. 相比之下, SvdA 方法为了追求更高的通用性, 在实验中并未使用缩放因子.

3.1 基线

除了使用 LoRA 与 VeRA 的实验结果作为本次实验的基线以外, SvdA 也将和以下基线进行测评指标的对比.

全参数微调 (FT): 将模型初始化为预训练模型的权重和偏差, 所有的参数都进行梯度更新. 下面实验中的基线采用先前工作中冻结其他层, 只去适应 GPT-2 模型最后两层的方法[35].

偏置项微调 (BitFit): 在训练时只更新偏置的值. 实验表明 BitFit 只会微调万分之 4 的参数, 它的微调比重略低于 LoRA 方法的比重[36].

适配器微调 (Adapters): 该方法在 Transformer 块之间插入两层适配器, 下面的实验涉及了多个不同类型的适配器. 由 Houlsby 等人[37]提出最早的适配器, 记为 AdptH. 其将适配器使用残差连接插入到注意力机制与前馈神经网络层之间. Lin 等人[38]提出只在多层感知机模块和 LayerNorm 之后各自插入一个适配器层的方法, 记为 AdptL. 类似的工作还有 Pfeiffer 等人[39]的工作中提到的方法, 记为 AdptP. 由 Rücklé等人[40]提出的基线, 通过动态移除部分适配器层来提高效率, 记为 AdptD.

主奇异值和向量微调 (PiSSA): 一种使用主奇异值和向量初始化 LoRA 低秩矩阵的方法. 下列实验中, 将秩等可训练参数相关超参设置与 LoRA 实验一致, 保留 Meng 等人[21]工作中的其他设置.

3.2 自然语言生成能力测评

实验采用了 GPT-2 模型[41], 在 E2E 数据集上进行微调, 并对微调后的自然语言生成能力进行了评估. 实验超参数基于 Hu 等人[5]的设置, 并加入了 Kopiczko 等人[22]关于 VeRA 的设定. SvdA 实验在一个 NVIDIA 4090 24 GB GPU 上完成, 通过超参数调优来确定学习率等超参数.

该实验使用了最后一个 epoch 的结果进行指标测试. 实验结果表明, SvdA 在保持与 VeRA 相近的可训练参数的同时, 在自然语言生成的多数指标的性能上优于 VeRA. 如表 1 所示, 带有星号 (*) 的方法的结果取自之前的工作. 与 LoRA 相比, SvdA 的可训练参数少了 3 倍以上; 与 VeRA 相比, SvdA 在多项指标上达到最优, 在实验中表现出更好的性能.

表 1 在 E2E 基准测试上不同微调方法的实验结果

方法	可训练参数量 (M)	BLUE	NIST	METEOR	ROUGE_L	CIDEr
FT*	354.92	68.2	8.62	46.2	71.0	2.47
AdptL*	0.37	66.3	8.41	45.0	69.8	2.40
AdptL*	11.09	68.9	8.71	46.1	71.3	2.47
AdptH*	11.09	67.3	8.5	46.0	70.7	2.44
LoRA*	0.35	68.9	8.69	46.4	71.3	2.51
PiSSA	0.35	68.5	8.62	46.2	70.8	2.44
VeRA*	**0.098**	70.0	8.81	46.6	**71.5**	2.50
SvdA	0.098	**70.1**	**8.82**	**46.7**	71.4	**2.54**

3.3 自然语言理解能力测评

下列实验采用 RoBERTa-base 模型, 在 GLUE 基准上评估了 SvdA 方法. 该基准测试涉及自然语言推断、文本蕴含、情感分析、语义相似等多个任务. 实验设置与 Hu 等人[5]的一致. 关于参数效率的实验与 SvdA 部分实验在一个 NVIDIA 4090 24 GB GPU 上进行.

实验采用了 CoLA 的马修相关性, STS-B 的皮尔逊相关性, 以及其他任务的准确性作为评价指标. 在所有情况下, 值越高表示性能越好. 实验结果如表 2 所示, 在自然语言理解能力的测评中, SvdA 相比拥有同量级可训练参数的 VeRA, 仍展现出了更为优秀的平均性能.

表 2 在 GLUE 基准测试上不同微调方法的实验结果

方法	可训练参数量 (M)	MNLI	SST-2	MRPC	CoLA	QNLI	RTE	STS-B	Avg
FT*	125	**87.6**	**94.8**	90.2	63.6	92.8	78.7	91.2	85.2
AdptL*	0.1	84.7	93.7	**92.7**	62.0	91.8	81.5	90.8	85.4
AdptL*	0.3	87.1	94.2	88.5	60.8	93.1	71.5	89.7	83.0
AdptH*	0.9	87.3	94.7	88.4	62.6	93.0	75.9	90.3	84.2
LoRA*	0.3	87.5	95.1	89.7	63.4	**93.3**	**86.6**	**91.5**	**86.6**
PiSSA	0.3	86.9	94.6	88.2	60.6	92.2	76.2	89.0	84.0
VeRA*	**0.031**	—	94.5	89.7	64.1	91.9	75.8	90.3	84.4
SvdA	0.036	85.5	93.8	90.0	**64.6**	92.0	77.6	90.9	84.9

为了更深入地了解不同微调方法对可训练参数利用效率, 实验评估了不同微调方法在 RTE 任务上的性能增益. 首先, 使用各个微调方法在 MNLI 任务上训练得到的模型作为基准模型, 然后在基准模型上进行 RTE 任务的微调. 最终, 使用微调后的模型与基准模型进行性能增益的比较. 性能增益计算公式如下:

$$\frac{accuracy_{\text{method}} - accuracy_{\text{baseline}}}{parameters_{\text{method}}} \times 100\% \tag{5}$$

其中, $accuracy_{\text{method}}$ 表示 method 微调方法在 MNLI 任务上微调 RoBERTa-base 模型得到的 RTE 指标准确率, 而 $accuracy_{\text{method}}$ 则表示将 MNLI 任务得到的模型, 针对 RTE 任务二次微调的 RTE 指标准确率. $parameters_{\text{method}}$ 表示微调方法的可训练参数量, 其计算方法为: 实际训练参数量/10^3. 实验结果如图 3 所示. 结果表明, SvdA 方法在每增加 1k 的可训练参数的情况下, 获得了最高的性能增益.

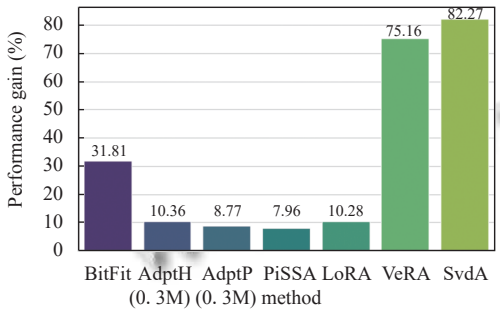

图 3 RoBERTa-base 模型在 RTE 任务上每增加 1k 参数量的性能增益对比

3.4 寻找秩的成本

由于 VeRA 的共享矩阵不需要额外的存储空间, 其 r 值不需要是低秩的. 在进行 VeRA 实验时, 选择一个适合的 r 值成为一个具有挑战性的问题. 而 SvdA 由于并没有使用低秩矩阵作为共享矩阵, 其不涉及秩的概念. 因此, 它能在单次实验中完成传统方法一组实验的工作量.

Kopiczko 等人[22]的研究工作中, 为了找到 LoRA 和 VeRA 在 RTE 任务上的最佳效果, 使用了两组秩并各自进行了多次实验. 尽管引入秩作为超参数在某种程度上提高了方法的灵活性, 但在不同任务中寻找一个适合的秩无疑会增加额外的工作量.

实验将 SvdA 与先前研究中表现更好的 VeRA 进行对比, 并观察 SvdA 方法是否能更高效地得到更为优质的实验结果. 对于 VeRA 的实验, 将使用 Kopiczko 等人[22]在 RoBERTa-base 实验中的学习率参数, 使用秩 r={1, 4, 16, 64, 256, 1024}进行多次实验, 并将结果与 SvdA 在 RTE 任务上的效果进行对比. 值得注意的是, SvdA 并没有秩这一超参数, 因此将 SvdA 实验单次结果作为与 VeRA 进行比较的基线.

当 rank 值为 1 024 时, 单张的 NVIDIA 4090 24 GHz GPU 无法满足 VeRA 实验的需求. 因此, 采用一张 NVIDIA A100 40 GHz GPU 来完成 rank 为 1 024 部分的实验. 实验结果显示, 相比于 VeRA 方法, 在 RTE 任务上, SvdA 方法表现更佳, 如图 4 所示. 当 VeRA 方法的 rank 值为 1 024 时, SvdA 方法以更少的实验成本取得了更优异的结果.

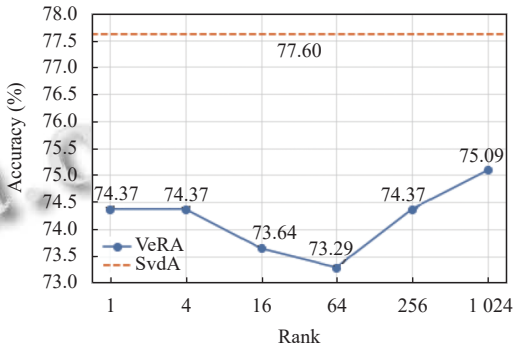

图 4 SvdA 和 VeRA 方法在 RTE 任务上的性能

4 结论与展望

本文提出了一种基于奇异值分解的适应微调方法 (SvdA), 旨在解决大语言模型在微调效率和存储成本方面的挑战. 通过将奇异值分解得到的对角矩阵和缩放向量作为可训练参数, SvdA 方法在多个自然语言处理任务中表现出色, 验证了其在减少参数数量和提高微调效率方面的有效性. 奇异值分解计算复杂且占用大量存储空间, 特别是对高维矩阵. SvdA 通过引入缩

放向量作为可训练参数, 利用共享酉矩阵有效地结合了奇异值分解方法, 提升了存储效率.

然而, 本文的研究也存在一些局限性. 例如, SvdA 方法在超大规模模型上的应用效果仍需进一步验证. 未来的研究工作将致力于以下几个方向: 首先, 从存储优化的角度进一步改进 SvdA 方法, 以提升其在更大规模模型和更广泛任务上的适应性和效率; 其次, 研究如何更有效地应用 SvdA 方法, 以应对大规模数据和模型的训练需求; 最后, 探索 SvdA 方法与其他微调技术的结合, 以实现更高效的模型微调.

参考文献

1 Zhao WX, Zhou K, Li J, *et al*. A survey of large language models. arXiv:2303.18223, 2023.

2 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. Nature Medicine, 2023, 29(8): 1930–1940. [doi: 10.1038/s41591-023-02448-8]

3 Brownlee J. How to avoid overfitting in deep learning neural networks. https://machinelearningmastery.com/introduction-to-regularization-to-reduce-overfitting-and-improve-generalization-error/. (2019-08-06).

4 Ding N, Qin YJ, Yang G, *et al*. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 2023, 5(3): 220–235. [doi: 10.1038/s42256-023-00626-4]

5 Hu EJ, Shen YL, Wallis P, *et al*. LoRA: Low-rank adaptation of large language models. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

6 Dettmers T, Pagnoni A, Holtzman A, *et al*. QLORA: Efficient finetuning of quantized LLMs. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 441.

7 Doering N, Gorlla C, Tuttle T, *et al*. Empirical analysis of efficient fine-tuning methods for large pre-trained language models. arXiv:2401.04051, 2024.

8 Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv:2303.15647, 2023.

9 Liu X, Zheng YN, Du ZX, *et al*. GPT understands, too. AI Open, 2023. [doi: 10.1016/j.aiopen.2023.08.012]

10 Liu X, Ji KX, Fu YC, *et al*. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 61–68.

11 Chavan A, Liu Z, Gupta D, *et al*. One-for-all: Generalized LoRA for parameter-efficient fine-tuning. arXiv:2306.07967, 2023.

12 Aghajanyan A, Gupta S, Zettlemoyer L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021. 7319–7328.

13 Sun XH, Ji YJ, Ma BC, *et al*. A comparative study between full-parameter and LoRA-based fine-tuning on Chinese instruction data for instruction following large language model. arXiv:2304.08109, 2023.

14 Zhang QR, Chen MS, Bukharin A, *et al*. Adaptive budget allocation for parameter-efficient fine-tuning. Proceedings of the 11th International Conference on Learning Representations. Kigali: OpenReview.net, 2023.

15 Zhang ZM, Ely G, Aeron S, *et al*. Novel methods for multilinear data completion and de-noising based on tensor-SVD. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014. 3842–3849.

16 Hastie T, Mazumder R, Lee JD, *et al*. Matrix completion and low-rank SVD via fast alternating least squares. The Journal of Machine Learning Research, 2015, 16(1): 3367–3402.

17 Feng X, Yu WJ, Li YH. Faster matrix completion using randomized SVD. Proceedings of the 30th IEEE International Conference on Tools with Artificial Intelligence. Volos: IEEE, 2018. 608–615.

18 Zhang J, Lei Q, Dhillon IS. Stabilizing gradients for deep neural networks via efficient SVD parameterization. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 5801–5809.

19 Yang HR, Tang MX, Wen W, *et al*. Learning low-rank deep neural networks via singular vector orthogonality regularization and singular value sparsification. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020. 2899–2908.

20 Chen YY, Tao QH, Tonin F, *et al*. Primal-attention: Self-attention through asymmetric kernel SVD in primal representation. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2024. 2840.

21 Meng FX, Wang ZH, Zhang MH. PiSSA: Principal singular values and singular vectors adaptation of large language models. arXiv:2404.02948, 2024.
22 Kopiczko DJ, Blankevoort T, Asano YM. VeRA: Vector-based random matrix adaptation. Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview.net, 2024.
23 Tao SY, Shen CY, Zhu L, *et al*. SVD-CNN: A convolutional neural network model with orthogonal constraints based on SVD for context-aware citation recommendation. Computational Intelligence and Neuroscience, 2020, 2020: 5343214.
24 Saxe AM, McClelland JL, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Proceedings of the 2nd International Conference on Learning Representations. Banff, 2014.
25 Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. Proceedings of the 7th International Conference on Learning Representations. New Orleans: OpenReview.net, 2019.
26 Sun YF, Zheng L, Deng WJ, *et al*. SVDNet for pedestrian retrieval. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 3820–3828.
27 Novikova J, Dušek O, Rieser V. The E2E dataset: New challenges for end-to-end generation. Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Saarbrücken: ACL, 2017. 201–206.
28 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002. 311–318.
29 Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Proceedings of the 2nd International Conference on Human Language Technology Research. San Francisco: Morgan Kaufmann Publishers, 2002. 138–145.
30 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the 2005 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor: ACL, 2005. 65–72.
31 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the 2004 Workshop on Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.
32 Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575.
33 Wang A, Singh A, Michael J, *et al*. GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels: ACL, 2018. 353–355.
34 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.
35 Li XL, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021. 4582–4597.
36 Zaken EB, Goldberg Y, Ravfogel S. BitFit: Simple parameter-efficient fine-tuning for Transformer-based masked language-models. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 1–9.
37 Houlsby N, Giurgiu A, Jastrzebski S, *et al*. Parameter-efficient transfer learning for NLP. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 2790–2799.
38 Lin ZJ, Madotto A, Fung P. Exploring versatile generative language model via parameter-efficient transfer learning. arXiv:2004.03829, 2020.
39 Pfeiffer J, Kamath A, Rücklé A, *et al*. AdapterFusion: Non-destructive task composition for transfer learning. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. ACL, 2021. 487–503.
40 Rücklé A, Geigle G, Glockner M, *et al*. AdapterDrop: On the efficiency of adapters in Transformers. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 7930–7946.
41 Radford A, Wu J, Child R, *et al*. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9.

(校对责编: 张重毅)